Genomic Signatures Of Adaptation In Eukaryotic Systems

Acuna Alonzo, Alin Patricia

Award date:
2020

Awarding institution:
University of Bath

GENOMIC SIGNATURES OF ADAPTATION IN EUKARYOTIC SYSTEMS

Alín Patricia Acuña-Alonzo

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

May 2019

Table of contents

TABLE OF CONTENTS **2**

ACKNOWLEDGEMENTS **6**

CONTRIBUTIONS **8**

ABSTRACT **10**

CHAPTER 1. INTRODUCTION **12**

MOLECULAR BASES OF MACROEVOLUTION 12

GENOMICS OF MAMMALIAN EVOLUTION 16

Mammalian Evolutionary Dynamics 16

Brain Evolution of Mammals and other taxa 22

Sexual selection 27

TRANSCRIPTIONAL SIGNALS OF ADAPTATION IN PLANTS 29

CHAPTER 2. BRAIN DEVELOPMENT RELATED GENE FAMILIES EXHIBIT SIZE VARIATIONS ASSOCIATED WITH NEURON AND GLIA CELL COMPOSITION AND ENCEPHALIZATION IN THE MAMMALIAN BRAIN **33**

ABSTRACT 33

INTRODUCTION 35

HYPOTHESES 38

MATERIAL AND METHODS 38

Cellular composition of the brain and encephalization data 38

Gene family annotations 40

Phylogenetic regressions of gene family size and cell composition parameters 41

Effect size 42

Power analysis 43

Effective number of tests 43
Effect size distribution randomisation test 44
Gene ontology term enrichment analysis 44
Co-expression and temporal gene expression analysis 45
Phylogenetic path analysis 46
RESULTS 47
DISCUSSION 51
TABLES 58
FIGURE LEGENDS 60
FIGURES 62
SUPPLEMENTARY MATERIAL 66
CHAPTER 3. SEXUAL SIZE DIMORPHISM IS ASSOCIATED WITH VARIATIONS IN GENE FAMILY SIZE IN GENE FAMILIES RELATED TO ORGANISM AND NEURONAL DEVELOPMENT 84
ABSTRACT 84
INTRODUCTION 86
HYPOTHESES 89
MATERIAL AND METHODS 89
Sexual size dimorphism and body mass. 89
Test for Rensch’s rule 90
Gene family annotations 90
Phylogenetic regressions of gene family size and cell composition parameters 91
Effect size 92
Power analysis 92
Effective number of tests 93
Effect size distribution randomisation test 93

Gene ontology term enrichment analysis 94
RESULTS 95
DISCUSSION 97
TABLES 101
FIGURE LEGENDS 103
FIGURES 104
SUPPLEMENTARY MATERIAL 109
CHAPTER 4. HERITABLE TRANSCRIPTOME SIGNATURES OF SOURCE POPULATION CLIMATIC CONDITIONS IN LAB-GROWN SALT TOLERANT *CAKILE MARITIMA* 111
ABSTRACT 111
INTRODUCTION 113
HYPOTHESES 115
MATERIAL AND METHODS 115
Illumina RNA sequencing 115
Transcriptome data annotation 116
Transcriptome profile clustering 117
Environmental data 117
Phylogenetic controlled correlations of gene expression and bioclimatic variables 118
Gene ontology term enrichment analysis 119
Variant calling 120
Phylogenetic tree construction 120
RESULTS 120
DISCUSSION 122
TABLES 127
FIGURE LEGENDS 128

FIGURES 129

SUPPLEMENTARY MATERIAL 133

CHAPTER 5. GENERAL CONCLUSIONS 139

REFERENCES 148

Acknowledgements

I would like express my profound gratitude to my main supervisor Dr. Araxi Urrutia for all her invaluable support and guidance, both on an academic and personal level, which were crucial to the fulfilment of this project.

I would like to thank Dr. Paula Kover, Prof. Jason Wolf, Dr Nicholas Priest, Dr. Nick Longrich and Dr. Humberto Gutierrez for all their support and helpful suggestions provided during the development of this work.

I also want to thank all the members (past and present) of the Laboratory 1.29 from the Biology Department and the Milner Centre for Evolution, for they support and friendship.

Also to The National Council on Science and Technology (Consejo Nacional de Ciencia y Tecnología, CONACYT) for the sponsorship provided for the development of this project.

To the Mexican community & friends in the UK that are an extended family on a different continent, specially Luz, Jorge M., Jimena, Atahualpa, Nik, Pola, Yan, Lupita, Pablo, Laura and Jeremy. The time spent with you was fundamental in keeping things in balance.

To Antonio, Amaya, Tabita, Celta, Dora, Iris, Jorge, Paola G.T. and Fer, for all this years of camaraderie and constant encouragment.

And last, but not least, my greatest heartfelt acknowledgement to my highest role models, my parents (Myrna and Victor) and brother (Victor) that are always present and supportive no matter what.

Contributions

The work presented in the thesis is my own work, produced with the help and advice from my supervisor Dr Araxi Urrutia as well as advice from other academics for specific chapters and contributions acknowledged below:

Additional academic supervision:

For chapter one, Dr Nicholas Longrich provided advice and suggestions on mammalian evolution in general.

For chapter two, Dr Alexandra de'Souza provided advice on brain cell composition and on brain evolution generally contributed neuron and glia number estimates for mammalian species used in Chapter two. Dr Humberto Gutierrez contributed general advice and proofread the revised version of the chapter. Dr Alejandro Gonzalez Voyer provided advice on phylogenetic path analysis. Dr Atahualpa Castillo Morales provided advice on many aspects, but particularly on discussions on random expectations calculated from random distributions and assessing significance of deviations from these.

For chapter three, Dr Serrano Meneses provided conceptual advice on sexual selection and on testing for the Rensch's rule. Prof. Tamas Szekely provided advice on size dimorphism as an estimate of sexual selection and the role of sexual selection on size in driving increases in species' body mass. Dr Atahualpa Castillo provided general advice, particularly on phylogenetic regression analysis.

For chapter four, Dr Paula Kover provided general advice on all aspects of the project as well as all transcriptome data used in the study.

Contributions:

For chapter two, Ms Karina Diaz and Paola Cornejo provided help and advice with gene expression and phylogenetic path analyses.

For chapter three, Dr Kathryn Maher, Mr Budd Nicholson and Mr Laurie Fabian contributed data on body mass for males and females with additional technical assistance from Mr Benjamin Padilla Morales.

For chapter four, Ms Kate Petty calculated the maximum likelihood phylogenetic tree for *Cakile maritima* populations used.

Abstract

Phenotypic variation between species or between individuals of a single species is encoded in their genomes, reflecting their specific evolutionary path. By comparing genomes and functional molecular data from several species and/or multiple individuals from a single species we can investigate the molecular signatures that underlie phenotypic diversity. The emergence and improvement of RNA and DNA sequencing and genome wide data analysis methods have facilitated the study of the molecular processes that shape phenotypic traits in an ever-expanding set of species beyond traditional model organisms. Taking advantage of these technologies and accumulating data, here I explore the molecular basis of two important phenotype traits, brain cell morphology and sexual selection in mammals and local adaptation in a plant species for which molecular underpinnings remain poorly understood. First, using a comparative genomics approach on 19 mammalian species for which brain cell composition data and fully sequenced genomes are available, I determined the relationship between variation in brain cellular composition of the brain and variations in gene family size (GFS). The results revealed significant associations between changes in gene family size and different cell composition parameters. Gene families associated neuron number, neuron density, glia to neuron ratio and encephalization were enriched various developmental related functions in the brain and generally including cell projection and neuron development whereas families associated with glia to neuron ratio were enriched in translation and cell migration. Immune system functions were also enriched among encephalization and neuron number associated gene families. These results are not explained by phylogenetic relatedness or associations among the different variables studied. Secondly, using a similar approach, I investigated the link between GFS changes and sexual size dimorphism (SSD) in mammals to elucidate the molecular signatures of differing degrees of sexual selection. From the 44 mammalian species assessed, it was observed that 729 gene families are significantly correlated with changes in GFS and SSD. Families expanding in line with increases in size dimorphism were found to be enriched in regulation of cell adhesion and stimulatory

C-type lectin receptor signalling pathway. Interestingly, gene families associated with reduced size dimorphism were found to be enriched in various aspects of brain development. These results could suggest that monogamous lineages experience higher selection on brain complexity to deal with the more elaborate social structures. Associations were not accounted for phylogenetic relatedness or by co-varying overall body mass. Finally, I examined transcriptome profiles of 19 samples of the searocket *Cakile maritima* (family Brassicaceae), stemming from five different locations. Using a phylogenetically corrected comparative approach it was shown that changes in gene expression are associated with various aspects of bioclimatic variation. These associated genes were significantly enriched in various functional categories related to the response mechanisms of plants to stress, e.g. regulation of endopeptidase activity, sugars metabolic processes, cell redox homeostasis, regulation of gene expression and DNA duplex unwinding. Crucially, these results are not explained by non-heritable phenotypic plasticity as transcriptome profiles were obtained from plants grown in controlled conditions from collected seeds. These results contribute to the understanding of the genetic mechanisms of adaptation to changing environments in plants. Overall, the findings presented in this thesis provide novel insights into the molecular background of complex phenotypes in eukaryotes using genomic and transcriptome data bioinformatics analyses. Importantly, the results presented here could not be inferred from single species analyses in model organisms as the phenotypes of interest. Large brains, high and low sexual selection and local adaptation to arid environments are not found in mammalian rodent models or the model plant *Arabidopsis thaliana*.

Chapter 1. Introduction

Molecular bases of macroevolution

Since life first appeared on Earth, there has been a massive expansion in the number of species and the phenotypic innovations that they exhibit. Macroevolution, refers to the study of evolutionary changes at or above the species level and usually the evolution of species over a geological time (thousands to million years) (Hlodan, 2007). In contrast, is considered that microevolutionary processes shape genetic diversity within a single species or a population, over a timescale of a few generations.

Is not easy to answer the question on whether macroevolution is the result of the gradual accumulation of evolutionary changes and there is an extended debate between punctuated equilibrium and gradualism theories (Erwin, 2000, Gould, 2002, Levinton, 2001). Emergent properties (revealed by discontinuities in the fossil record and developmental innovations) could imply that macroevolution does not result merely from the accumulation of microevolutionary processes (Erwin, 2000).

However, the debate is complicated because macro and micro evolution are studied from different perspectives and using different methods (palaeontology, phylogenetics and evo-devo in the case of Macroevolution; population genetics in the case of Microevolution). The implementation of models that can account for evolutionary change across a broad range of timescales should help to understand evolutionary pattern and process (Uyeda et al., 2011).

Because there is a poor understanding of the relationship between morphological and genetic divergence among distantly related taxa some authors advice that it is not useful to distinguish sharply between microevolution and macroevolution (Levinton, 2001).

Instead, they propose a different definition focused on character-state differences rather that the taxonomic level or time scale (Levinton, 1983). For Levinton (2001), macroevolution is “the sum of processes that explain the character-state transitions that diagnose evolutionary differences of major taxonomic rank."

If we accept that the accumulation of evolutionary changes can have an effect in the evolution of species in a geological time scale, then we should consider mutation, natural selection, genetic drift and gene flow as potential factors in macroevolution.

The primary source of variation is mutation, defined as any change in the base sequence of the DNA (Smith, 1998). Mutations may involve changes of a single base (substitution, insertion/deletion) or changes of sections of the DNA (inversion, duplication/ deletion). As the only process that generates new alleles, mutation is the ultimate source of genetic diversity (Jobling, 2014). It occurs randomly and it is heritable only when it acts on the DNA of germ cells. Most of the mutations that alter the fitness of organisms will lower it (Smith, 1998).

Natural selection is the process by which allele frequencies change due to differences in the survival and reproduction of individuals with different genotypes. Selection acts differentially along the genome because it only affects the region surrounding the gene that is under selection pressures. Some mutations in functionally important sequences confer a fitness benefit, hence, are subject to positive selection (Strachan and Read, 2010). These mutations are the basis of adaptive evolution (Charlesworth, 2012). In contrast, deleterious mutations are selectively eliminated by purifying (negative) selection. Sequences under purifying selection appear to be strongly conserved and subject to evolutionary constraint (Strachan and Read, 2010).

Genetic drift is the fluctuation in allele frequencies in a finite population due to random variations in the contribution of each individual to the next generation (Jobling, 2014). It acts on the existing genome variation and it is dependent on the size of population. The smaller the population, the more drastic the shifts in allele frequencies. Population bottlenecks and founder effects can lead to genetic drift processes. Through genetic drift neutral mutations can become fixed in the population (Strachan and Read, 2010).

Gene flow refers to the exchange of alleles from one population to another. Gene flow increases the genetic diversity within the population (Templeton, 2006).

With the availability of complete genomes of hundreds of species, it is possible now to compare gene families over different evolutionary timescales (close and distant species).

One of the most relevant genomic processes shaping functional innovation is the duplication of genes. Gene duplication is recognized as a significant factor in the evolution of genes, genomes, and species (Taylor and Raes, 2005). Particularly, it has been considered as key contributor to phenotypic diversity as a source of gene novelty (Conrad and Antonarakis, 2007, Magadum et al., 2013). The mechanisms that can give rise to gene duplication are unequal crossing over, retroposition, and chromosomal or genome duplication (Zhang, 2003). Unequal crossing over occurs at meiosis when the homologous chromosomes do not pair correctly and it results in an uneven exchange of material. This mechanism generates tandem gene duplications and plays an important role in the origin and expansion of gene families (Taylor and Raes, 2005). Retroposition results by the reverse transcription of mRNA into cDNA and its subsequent insertion in the genome. Chromosomal or genome duplication result from the non-disjunction among homologous chromosomes during meiosis (Zhang, 2003).

The retention of duplicated genes in the genome depends on different factors, such as function, mode of duplication, expression rate and species (Taylor and Raes, 2004). Gene duplication might result in nonfunctionalization (one copy losses its function by pseudogenization), neofunctionalization (one copy acquires a novel adaptive function while the other retains the original function) or subfunctionalization (both genes specialize to perform complementary functions) (Conrad and Antonarakis, 2007).

Gene duplication contributes to organismal evolution by providing new genetic material for mutation, drift and selection to act upon, and resulting in specialized or new gene functions (Zhang, 2003). An important product of gene duplication and divergence are gene families (Taylor and Raes, 2005). Gene families are groups of homologous genes with similar functions (Demuth et al., 2006). Among species, the process of differential gene gain and loss result in gene families of different sizes (Demuth et al., 2006). Various gene families essential for organismal fitness have been described (e.g. inmunoglobulins, hemoglobins, heat shock proteins, histocompatibility antigens) (Taylor and Raes, 2005). Moreover, it is proposed that changes in gene family size (GFS) are important in evolution (Demuth et al., 2006, Chen et al., 2010). For example, the olfactory receptor (OR) family, that constitute the largest mammalian gene family (>1,000 genes), is considered to have evolved under different selection pressures in humans and great apes (Gilad et al., 2003). Comparing the gene sequences of OR genes between humans, chimpanzees and orangutan, Gilad *et al.* (2003) found that the different species have distinct patterns of variability in those genes. Particularly, the study suggests that chimpanzee intact OR genes have undergone purifying selection, whereas human OR genes underwent positive selection. In both species OR pseudogenes appear to evolve under no evolutionary constraint. From this results, it is proposed that the observations reflect the distinct lifestyles of the species that led to different sensory requirements (Gilad et al., 2003).

Genomics of mammalian evolution

Mammalian Evolutionary Dynamics

Fossil and molecular studies are essential to establish origins, relationships and diversification of the organisms (Schrago et al., 2013). Technological improvements for this kind of analyses, in addition to the increase in new localities and specimens analyzed, have led to the current resurgence of interest in mammalian evolution (Archibald, 2006). It has been described that mammals were already widespread and ecologically diverse by the middle of the Jurassic period, about 165 million years ago (Ma) (Cifelli and Davis, 2013). However, issues relating to the evolutionary affinities of fossils (Schrago et al., 2013) and some basal branches of the family tree among this taxa are still on debate (Cifelli and Davis, 2013).

In general, according to the fossil record it is proposed that the story of the origin and evolution of mammals falls into three distinct phases (Kemp, 2005):

1) The lineage of amniotes that led to the origin of mammals (appeared about 305 Ma).
2) The Mesozoic mammals. Where numerous subgroups radiated from the mammalian ancestor.
3) The disappearance of most species of larger mammals from the fossil record and the shaping of the modern fauna (10-20 thousand years ago).

Mammalian origins

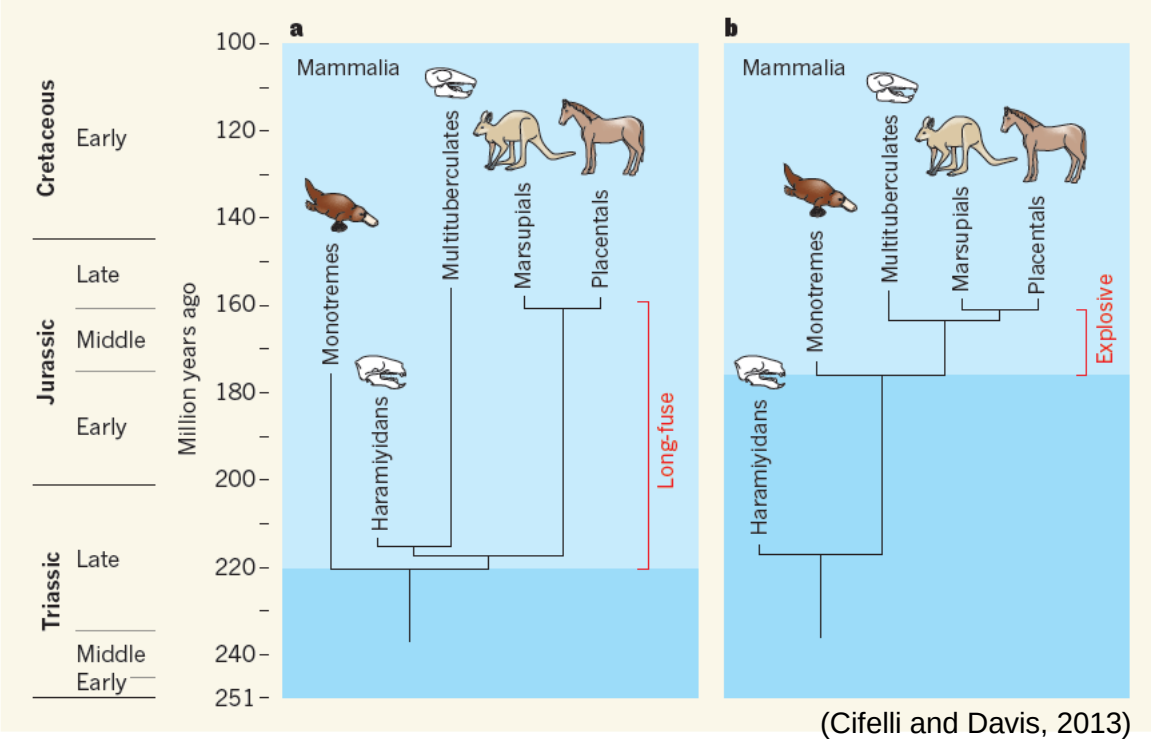
Regarding the origin of this taxa, it is recognized that the inclusion (or not) of certain early forms represents “a major sticking point” (Cifelli and Davis, 2013). For instance, disagreement remains over the taxonomy and phylogeny concerning Allotheria (Multituberculata and Haramiyida) (Bi et al.,

2014, Butler, 2000) and the resulting implications on the origin of mammals. A study by Zheng *et al.* (2013) described a new haramiyid from the Jurassic period, and the phylogenetic analysis placed Haramiyida within crown Mammalia. Suggesting the origin of crown Mammalia in the Late Triassic period (before the break-up of Pangaea) and its diversification in the Jurassic.

More recently, a research conducted by Bi and collaborators (2014) reported three new species of a new clade (Euharamiyida) from near-complete specimens. Phylogenetic analysis recognized it as the sister group of Multituberculata and placed Allotheria within the Mammalia supporting a Late Triassic origin of mammals approximately 208 Ma (Bi et al., 2014).

In contrast, a study carried out by Zhou *et al.*(2013) proposes that haramiyidans are separated from multituberculates by incorporating the observed features of a new fossil from the Middle Jurassic (Megaconus). It suggests that haramiyidans constitute a mammaliaform clade outside Mammalia and estimates their origin in the Middle Jurassic (about 175 Ma). Descriptions of this fossil showed specialization for masticating plants (herbivorism) and fur presence, indicating that this characteristics were already present at that stage (Zhou et al., 2013).

The two alternative interpretations of early mammalian history dependent on the inclusion of the haramiyidans as mammals are summarized in Figure 1.



(Cifelli and Davis, 2013)

FIGURE 1. Alternative proposed interpretations of early mammalian history. **a)** Inclusion of haramiyidans within Mammalia suggesting a Late Triassic origin for mammals. **b)** Placement of haramiyidans outside Mammalia implying a Middle Jurassic origin.

Mammalian radiation

While dinosaurs dominated the terrestrial fauna in the Mesozoic, numerous animal subgroups radiated and diversified from the mammalian ancestor (Kemp, 2005). Those early mammals that coexisted with dinosaurs in the Mesozoic are of great interest to understand major issues in phylogeny and early evolution of the group (Meng, 2014). It is described that a “good deal of evolution” occurred in this Era, specially dental evolution (origin of tribosphenic tooth) and the roots of marsupials and placentals (the two major modern taxas) (Kemp, 2005).

The recent discovery of a complete and well-preserved skull of a gondwanatherian (from Upper Cretaceous in Madagascar), the species *Vintana sertichi*, contributes on this issues and reveals further morphological

diversity among early mammals (Krause et al., 2014). The cranium exhibits mosaicism of primitive and derived features, and the research authors attribute the acquisition of this morphology to the years of evolution in geographic isolation (Krause et al., 2014). This study places Gondwanatheria within Allotheria, supporting the Allotheria grouping (ancient origin for Mammalia), and furthermore, supports the idea that Pangaea's fragmentation contributed the radiation during the Mesozoic (Weil, 2014).

Even though the Mesozoic diversification of the mammalian groups is not the same as the ones we see today (Weil, 2014), basal diversifications of extant groups occurred in this Era (Luo, 2007).

Extant mammalian ancestors

The end of the Mesozoic is characterized by the mass extinction of dinosaurs along with many other taxa (Kemp, 2005). At the beginning of Tertiary, the surviving mammalian lineages underwent great diversification, both on the ecologic and taxonomic senses (Cifelli et al., 2004). For much of this Period, independent evolutionary radiations were occurring simultaneously on different areas and, for the first time, mammals of middle to large body size originated (Kemp, 2005). The evolution of the earliest mammals is known to have occurred in successive diversifications, however, the dates of today´s mammals ancestors evolution is still debated (Luo, 2007).

Extant mammals are classified into three main groups: monotremes, marsupials and placentals (Cifelli and Gordon, 2007). Marsupials and placentals represent 99% of all living mammal species, showing great ecomorphological diversity (Luo, 2007). Placentals is the most varied and diverse group (Cifelli and Gordon, 2007), hence, have been broadly investigated. Various studies have proposed alternative models for the placental diversification (Archibald and Deutschman, 2001):

- “Explosive” → divergence of most superorders and orders happened near and following the Cretaceous – Tertiary (K/T) boundary. Palaeontological evidence tends to support this model (Archibald and Deutschman, 2001, Bininda-Emonds et al., 2007).
- “Long fuse” → earlier diversification of superorders.
- “Short fuse” → diversification of both superorders and orders well back in the Cretaceous. (Cifelli and Gordon, 2007).

Molecular studies

In recent decades, the comparative and evolutionary analysis of molecular data has greatly contributed to reconstruct phylogenetic relationships (San Mauro and Agorreta, 2010). Particularly, the rapid growth in the availability of molecular data has increased the use of molecular-clocks leading to “a profound effect on our understanding of the temporal diversification of species and genomes” (Kumar, 2005). The molecular-clock hypothesis relies on the idea that molecular evolution occurs at an approximately uniform rate over time (Kumar, 2005).

It has been observed that molecular and fossil data can give dissimilar views of the evolutionary past (Bininda-Emonds et al., 2007). Generally, molecular studies, while highly variable, estimate divergence times much older than the ones based on fossils (Dos Reis et al., 2012, Goswami, 2012). Molecular analysis agree on estimates of a Cretaceous origin for Placentalia, although they differ on the diversification patterns relative to the K/T event (Dos Reis et al., 2012). For instance, a research carried out by Bininda-Emonds and collaborators (2007) suggested a delayed-rise scenario for the temporal pattern of extant mammalian diversification. Their analysis results of supertree algorithms including nearly all extant mammals, showed low diversification rates across the K/T boundary and a subsequent rise throughout the Eocene and Oligocene (56-65 Ma).

It is proposed that early molecular studies had methodological issues (e.g. inadequate dating methods and simplistic interpretations of the fossil record) (Dos Reis et al., 2012). However, in recent decades molecular divergence studies are implementing sophisticated analyses of entire genomes (Goswami, 2012) and Bayesian dating methods in order to circumvent those problems (Dos Reis et al., 2012). Hence, the gap between molecular and palaeontological estimated divergences is being reduced.

The construction and analysis of a molecular supermatrix (with likelihood-based methods and relaxed molecular clocks) produced divergence time estimates largely consistent with the fossil record, showing intraordinal divergences near the K/T boundary (Meredith et al., 2011). These results are consistent with the hypothesis that the KT mass extinction played an important role in the early and adaptative diversification of mammals (Meredith et al., 2011).

More recently, implementing a Bayesian method, the analysis of 36 nuclear and 274 mitochondrial genomes produced a date of origin for present-day placental orders (except Primates and Xenarthra) between 45 and 65 Ma (Dos Reis et al., 2012). This estimate shows high convergence with the fossil record that proposes that most placental mammal orders diversified after the K/T mass extinction (Goswami, 2012).

Thus, recent molecular studies are beginning to improve the congruence between molecular and palaeontological data (Goswami, 2012). Both types of data are useful and necessary in systematics; constitute independent and complementary sources of information that allow to cross-validate hypotheses regarding the evolutionary history and dynamics of mammals (San Mauro and Agorreta, 2010).

Brain Evolution of Mammals and other taxa

Evolution of the vertebrate brain

At the present, we observe a large amount of phenotypic variation among vertebrates (Striedter, 2005). One of the traits that shows marked variation across this taxa is brain size (Shultz and Dunbar, 2010, DeFelipe, 2011). The brain is the most-studied vertebrate organ and has been of great interest for its important role in evolution since the late nineteenth century (Caspari, 1963, Jerison, 1973, Jensen et al., 1979, Sylvester et al., 2010, Roth and Wullimann, 2001). Among vertebrates, it is observed that encephalization (brain size relative to body size) has often increased and it is suggested that these increments were biologically significant since they occurred early on in the evolution of various major lineages (mammals, birds and cartilaginous fishes) (Striedter, 2005). Furthermore, it is assumed that the vertebrate brain did not evolve through a linear progression but in at least four distinct radiations (Roth and Wullimann, 2001), and between and within those, brain size varies in an orderly manner (Figure 2) (Northcutt, 2002). The four main radiations are:

- Agnathans (jawless vertebrates). Generally with the smallest brains relative to body size (Northcutt, 2002).
- Chondrichthyes (cartilaginous fishes). Many of them have brains as large for their body size as those of birds and mammals (Northcutt, 2002).
- Osteichthyes (ray-fined fishes or bony fishes). Most members of this group have considerably larger brains for the same body size than Agnathans (Northcutt, 2002).

- Tetrapods (anamniotes [amphibians] and amniotes [reptiles, birds and mammals]). Reptile brains are approximately two to three times larger than the brains of most amphibians of the same body size. Brains of birds and mammals have are between 6 to 10 times larger than the brains of reptiles of the same body size (Northcutt, 2002).

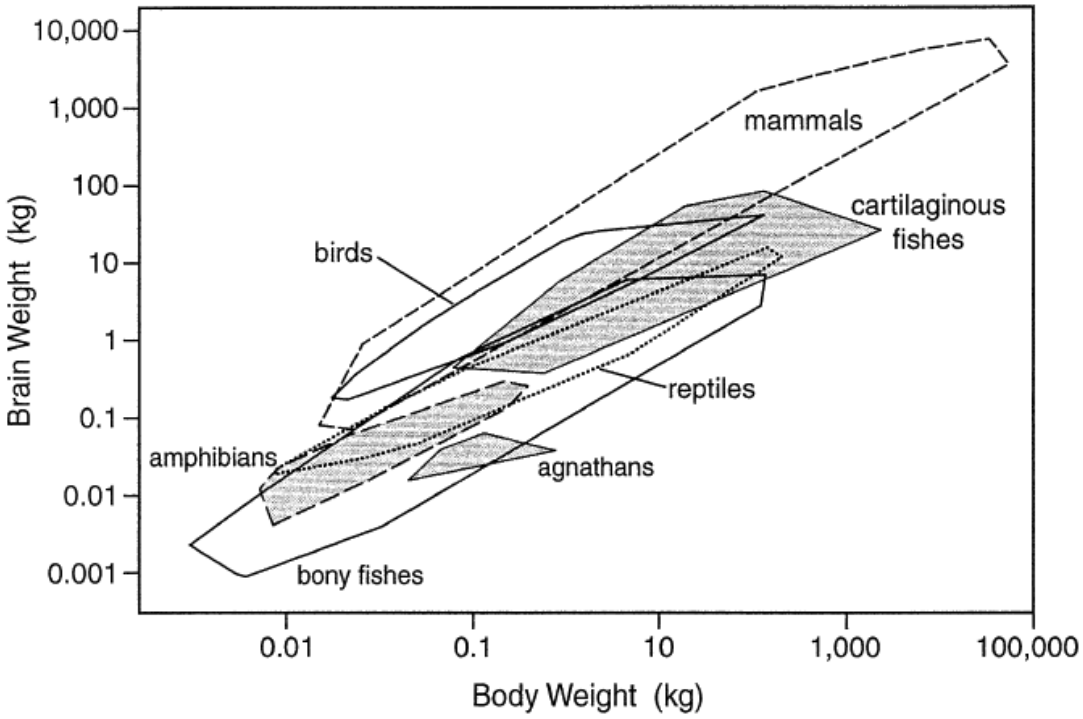


FIGURE 2. Vertebrate brain weights plotted against body weights. Data is expressed as minimal convex polygons for each of the major groups (Northcutt, 2002).

Most of the major vertebrate groups show a ten-fold range in brain size and is observed that brain size increases with body size in an allometric (non proportional) manner (Northcutt, 2002). The allometric slopes are estimated from 0.21 (agnathans) to 0.74 (mammals) (Northcutt, 2002).

Generally, in the history of diversity in encephalization, the evolution of the brain is considered as “conservative” (Jerison, 1979). Based on the observation that, within an adaptive zone, when a specific grade of

encephalization was achieved, the given taxon tended to conserve it. Only few exceptions, like mammals, are proposed to have had continued progressive evolution within an adaptive zone ("specialized in encephalization") (Jerison, 1979).

Closely related species tend to have more similar brains than distant relatives ones (Striedter, 2005). Specifically, across living vertebrates the major brain divisions are conserved, indicating that much of their organization arose with the origin of vertebrates or shortly later (Northcutt, 2002). It has been observed that changes in the internal organization of brains tend to occur along with their variation in size (Striedter, 2005).

Mammalian brain evolution

Across mammals, although evolutionary encephalization is assumed to be a general trend (Jerison, 2007, Shultz and Dunbar, 2010), analyses conducted with fossil data testing for temporal changes in relative brain size over time, have shown that there is wide variation across groups in this taxa (Shultz and Dunbar, 2010). In general, analysis show that the early mammals had lower measures of encephalization than their descendants (Jerison, 1979, Striedter, 2005), which leaded to suggest that is a reflection of the "more diverse adaptive modes in later taxa" (Jerison, 1979).

As showed on Figure 3, primates and cetaceans have the largest rates of relative brain size (Striedter, 2005). On the other hand, non-placental mammals, insectivores, marsupials and rodents have the smaller measures (Northcutt, 2002). It is proposed that much of this variation may be due to developmental constraints with one or more selective pressures been operative (Northcutt, 2002).

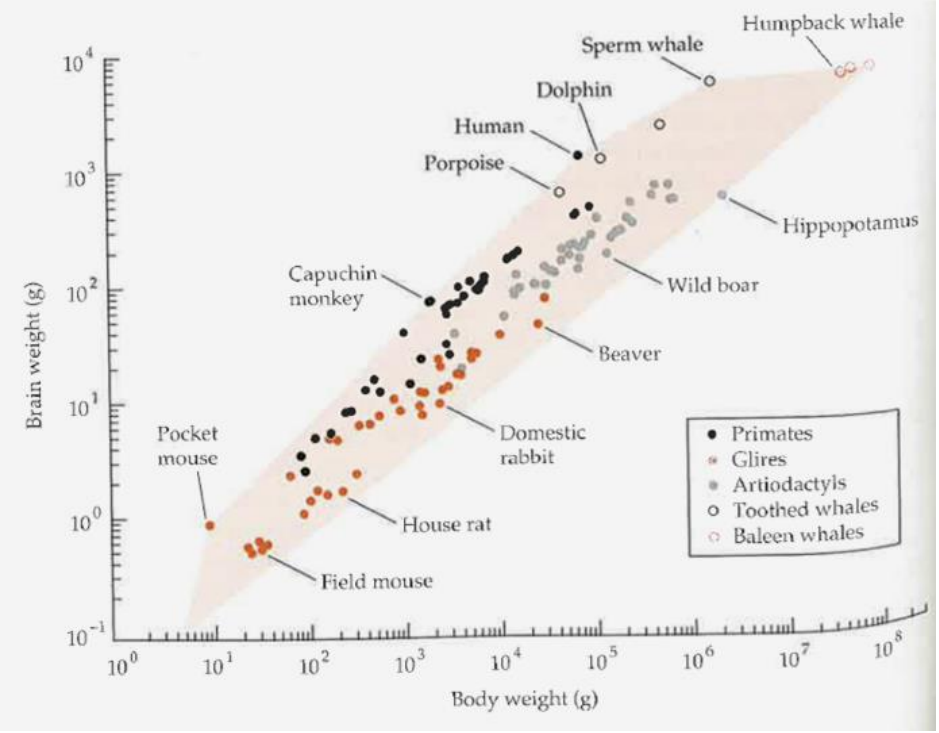


FIGURE 3. Relative brain size variation among mammals. Brain weights plotted against body weights (Striedter, 2005) .

It has been observed that between groups with macroevolutionary increase in brain size, primates have the highest encephalization slope (Shultz and Dunbar, 2010). Particularly, the hominid lineage has undergone a dramatic increase of endocranial volume over the past three million years (Lefebvre, 2012, Roth and Dicke, 2005). Analysis carried out with measures for relative brain size show that humans have the largest brains (Striedter, 2005). The human brain is about 3 times larger than the one of the chimpanzee, our closest living relative (Navarrete et al., 2011), even though adults of both species have similar body size measures (Striedter, 2005). Since brains are metabolically costly, and is considered that selective pressures act on structures to not surpass a “minimum essential size”, it is proposed that this enlargement has a functional basis (Shultz and Dunbar, 2010, Jason A. Kaufman et al., 2003, Clark et al., 2001). Moreover, as species are expected to maximize the cost benefit ratio of supporting expensive tissues, is

suggested that encephalization is associated with evolutionary advantages (Shultz and Dunbar, 2010).

Various theories have been proposed attempting to elucidate the way that metabolic requirements of large brains are offset. For instance, the expensive-tissue hypothesis suggests that is related to a corresponding reduction of the digestive tract (Aiello and Wheeler, 1995). This reduction was possibly allowed by an improvement in diet quality and feeding behavior (Roth and Dicke, 2005, Fish and Lockwood, 2003). By contrast, Navarrete *et al.* (2011) hypothesize that human encephalization was made possible by a combination of stabilization of energy inputs and a redirection of energy from locomotion, growth and reproduction. More recently, it has been proposed that the high costs of the development of human brain require a compensatory slowing of body growth rate (Kuzawa et al., 2014). Ultimately, brain evolution appears to be influenced by multiple factors including selection on behavioral characteristics (Schoenemann, 2013), physiological constraints and foraging strategies (Fish and Lockwood, 2003, Striedter, 2005).

Genes and Brain Size Evolution

Some genetic studies have been carried out to identify genes that affect brain features. In particular, the group of microcephaly genes is of great interest to study brain size evolution (Montgomery et al., 2011, Cox et al., 2006). Primary microcephaly is an autosomal recessive neurodevelopmental disorder in which gene mutations cause severe reduction in brain growth (Cox et al., 2006, Evans et al., 2004). Because affected individuals present a brain size that is similar to that of early hominids, its suggested that “these genes might have had a role in evolutionary expansion of the primate brain” (Cox et al., 2006). In a research carried out across anthropoid primates, Montgomery *et al.* (Montgomery et al., 2011) analyzed four genes associated with microcephaly (ASPM, CDK5RAP2, CENPJ, MCPH1) and suggested that adaptative evolution on two of them (ASPM and CDK5RAP2) has been

involved in the evolution of brain size. Earlier, Evans and collaborators (Evans et al., 2004) showed that the *microcephalin* gene in hominid lineages presented accelerated evolution which is coupled with signatures of positive selection during primate evolution, thereby suggesting a contribution to increased brain size in primates (Evans et al., 2004).

Recently, using a genome-wide comparative approach between 39 mammalian species, it was shown that variations in gene family size are associated with increased encephalization in mammals. It is suggested that this relationship reflect an evolutionary response to the functional demands underlying brain expansion (Castillo-Morales et al., 2014). This study further revealed that the main biological functions with higher enrichment among encephalization-associated gene families are cell signaling, immune regulation and chemotaxis (Castillo-Morales et al., 2014).

The increase of comparative genomics and related technologies has improved the study of human evolution, however, our understanding of the relationship between genetic and phenotypic changes remains inconclusive (Preuss, 2012). Furthermore, genome-wide determinants underlying mammalian encephalization are largely unknown.

Sexual selection

A considerable amount of phenotypic diversity results from variations in sexual fitness (i.e., sexual selection) (Andersson, 1994). The concept of sexual selection was defined by Darwin as depending “on the advantage which certain individuals have over other individuals of the same sex and species, in exclusive relation to reproduction” (Darwin, 1901). Thus, sexual selection is suggested to be distinguished from natural selection as it results from differences in mating success (Hosken and House, 2011). Furthermore, the effects of sexual and natural selection on the same characteristics can be

opposed. For example, conspicuous ornamentation or coloration have been observed to provide sexual selection advantage but at the same time increase the risk of attack by predators (Hernandez-Jimenez and Rios-Cardenas, 2012, Morgans et al., 2014, Husak et al., 2006, Håstad et al., 2005).

Sexual size dimorphism

Sexual selection can drive the evolution of dimorphism between males and females in morphological, behavioural and physiological traits (McPherson and Chenoweth, 2012). Body size dimorphism, is a most conspicuous feature in many animals, and it is thought to be caused by sexual selection (Andersson, 1994). Sexual size dimorphism (SSD) refers to the morphological differentiation in body size of sexually mature males and females within a species (Fairbairn, 1997). These differences prevalent among animals, are considered to be adaptive and with emergence occurring at different developmental stages (Dos Remedios et al., 2015, Klenovšek and Kryštufek, 2013). In taxa with larger males, generally SSD is observed to increase with body size, conversely, when females are larger SSD decreases. This trend, denominated Rensch's rule, is mainly explained as a result of sexual selection in males and energetic constraints that reproduction entails in larger females (Noonan et al., 2016, Rudoy and Ribera, 2017). SSD is of great evolutionary importance in determining a number of morphological and behavioural traits as well as long term species viability. Various studies have revealed patterns of SSD and mechanisms that originate them (review in (Fairbairn et al., 2007)), however, the genomic signatures that underlie them remain largely unexplored.

Transcriptional signals of adaptation in plants

The capacity of adaptation to local conditions in plants is a process that impacts their ability to thrive across broad geographic ranges (Halbritter et al., 2018, Leimu and Fischer, 2008). Similarly, genetic adaptation and adaptive phenotypic plasticity influence how populations persist under changing conditions, including climate change (Nunney, 2015, Hoffmann and Sgrò, 2011). Factors like population size, gene flow, heritable variation, and the rate of environmental change affect this ability to evolve local adaptation (Williams et al., 2008, Leimu and Fischer, 2008).

As sessile organisms, plants are easily influenced by their surroundings, therefore, during evolution have developed molecular and cellular mechanisms to respond to the changing environments over time (Shao et al., 2007, Sham et al., 2015). For instance, several studies on genetic adaptation along elevation gradients indicate that this is a widespread phenomenon in plants (Halbritter et al., 2018).Recently, Lobréaux *et al.* (2019a) revealed a link between differences in elevation and genes related to stress response in the plant *Arabis alpina* (Brassicaceae family) through a genome wide analysis. Thus, supporting that genetic diversity along the gradient is shaped by environmental pressures, leading to local adaptations (Lobréaux and Miquel, 2019a). Similarly, a study carried out on plants of the species *Arabidopsis lyrata* growing on different soil types, allowed to identify specific polymorphisms associated with environmental conditions (Turner et al., 2010). By sequencing the pooled DNA of plants from serpentine and granitic soils it was observed that the polymorphisms more strongly associated with soil type are enriched in gene ontology terms of metal ion transmembrane transporter activity and calcium ion binding. Thus, suggesting that polymorphisms are differentiated as a result of local adaptation to soil type.

Regulation of gene expression is considered an essential strategy for the adaptation to environmental stress conditions influenced by biotic and abiotic stress factors (Sham et al., 2015).

Transcriptome analysis on diverse plant populations covering varied geographical and environmental gradients are emerging as a useful approach to enhance the understanding of environmental factors and genetic mechanisms underlying locally adaptive trait variation (Akman et al., 2016). The current improvements in next-generation sequencing have greatly benefited this field of research (Schoville et al., 2012, Pespeni et al., 2013, Yang et al., 2015, Akman et al., 2016).

Although the majority of studies carried out to investigate the plant mechanisms in response to environmental changes have used the model organism *Arabidopsis thaliana* (Wu et al., 2012), recent research on other species has allowed the discovery of different mechanisms (Wu et al., 2012, Wong et al., 2005, Bressan et al., 2001). For instance, the close relative *Thellungiella* has been studied due to its intrinsic tolerance to abiotic stress including salinity, cold, drought and oxidative stresses (Amtmann, 2009). Through comparing the genomes of *Thellungiella salsuginea* and *Arabidopsis thaliana*, it was observed that stress related gene families in *T. salsuginea* underwent selective expansions during evolution (Wu et al., 2012). Moreover, it is proposed that it is the result of various duplication events. Wu and collaborators found 21 transcription factor gene families expanded in the *T. salsuginea* genome, some of them containing members related with stress resistance, thus suggesting an association with the adaptation to extreme environments.

Divergent responses to salinity stress between *T. salsuginea* and *Arabidopsis thaliana* have been observed trough transcriptome and metabolite analysis (Gong et al., 2005). Under increased salinity, the upregulated genes only in *Arabidopsis* are mainly related to general defence and protein synthesis, which suggest that this allows a response of “global defence” based on injury (Gong et al., 2005). On the other hand, the upregulated genes in *Thellungiella* are related to protein folding, post-translational modification and redistribution, suggesting a strategy more of

resources and energy conservation. Interestingly, *Thellungiella* shows high pre-stress concentrations of various compounds that have been shown to have protective functions in osmotic imbalances. These results led to propose that *Thellungiella* has a prior preparation to stress (Gong et al., 2005). A study that analysed expressed sequence tags (ESTs) from cDNA libraries of stress-induced *Thellungiella* plants observed that gene expression in response to salinity has little overlap with the cold- and drought-induced responses (Wong et al., 2005). Moreover, the comparison of ESTs between the *Thellungiella* ecotypes Yukon and Shandong revealed some differences in their expression profiles that might reflect divergent physiological stress responses (Wong et al., 2005). For example, the EST corresponding to a lipid transfer protein (LTP4) showed contrasting levels of over-expression between the two ecotypes under salinity and drought stress. Interestingly, this protein appears to be involved in the production of epicuticular waxes that are more abundant in the *Thellungiella* leaves than in *Arabidopsis* (Wong et al., 2005). Many of the genes differentially expressed under abiotic stress treatments in *Thellungiella* correspond to proteins that have no known function. As an extension of this study Wong and collaborators (2006) constructed a cDNA microarray from the previously obtained stress-induced libraries of *Thellungiella* to examine its molecular response subjected to cold, salt, drought and drought followed by rewatering stress on a slow and prolonged manner. From the 154 transcripts differentially regulated in response to the different conditions there was little overlap among the transcripts associated to each stress, in agreement with the idea that *Thellungiella* displays distinct responses under different stresses (Wong et al., 2005).

A recent analysis of mRNA-seq data combined with common garden experiments on 19 phenotypic and ecologically diverse populations of the species *Protea repens* found an association between source population climate and gene expression patterns reflecting trait variation along environmental gradients (Akman et al., 2016). Moreover, these correlations are suggested to indicate population differentiation as result of heritable

changes in gene expression. One more example is the analysis of gene expression dynamics carried out on the shrub *Artemisia sphaerocephala* in response to different stresses (Zhang et al., 2016). This study revealed a set of transcripts involved in heat, cold, salt and drought responses, thus shedding light on the possible molecular adaptive mechanisms of this plant to desertic environments.

Chapter 2. Brain development related gene families exhibit size variations associated with neuron and glia cell composition and encephalization in the mammalian brain

Abstract

Brain size, morphology and composition are key to the evolution of cognition in mammals. Changes in cell proliferation and death, connectivity and the ratio of neurons to supporting glia cells have been proposed as major determinants of brain function and targets of selection but the genomic signatures associated with changes in brain cell composition have not been explored. Encephalization, brain size corrected by body mass, has been associated with variations in gene family size (GFS), which reflect changes in the relative relevance of the molecular functions they represent. Using a phylogenetically corrected comparative genomics approach, here we test whether changes in brain cell composition parameters are associated with GFS and whether such associations explain previously reported links between encephalization and GFS. We found a significant expansion of gene families in line with encephalization and neuron number beyond random expectations. Gene families associated with neuron density were enriched in cell projection and neuron development functions; families associated with glia to neuron ratio were enriched in translation and cell migration-related functions, while neuron number associated gene families were enriched in immune system functions. These associations are independent from previous associations between immune related gene families and encephalization, which were confirmed. Overall, these results provide the first insights into the molecular signatures of the evolution of cellular composition in the brain and suggest that gene family expansion has played a role in the evolution of brain size and composition in mammals.

Key words: genome evolution, glia to neuron ratio, neuron density, neuron number, gene family size

Main points

- We report a novel link between neuron number, glia to neuron ratio and neuron density with variations in gene family size in mammalian species.

- The associations of gene family size and neuron number, glia to neuron ratio and neuron density are not explained by encephalization or phylogenetic relatedness.

- Neuron number associated gene families are enriched in functional categories related immune system whereas neuron density and glia to neuron ratio are associated to the expansion of brain development-related gene families.

- This is the first analysis exploring the genomic signatures of brain cellular composition.

Introduction

Mammalian brains show marked variations in brain size, morphology and cellular composition. Absolute brain size can vary by up to ten fold between mammalian species with differences being associated with cognitive capacity, social group sizes and other behavioural indexes (Shultz and Dunbar, 2010, DeFelipe, 2011, Gibson et al., 2001). Given that species with larger brains also are larger, brain size is often corrected by overall body mass, a measure known as encephalization quotient (Jerison, 2012).

The remarkable expansion of the brain is considered a significant process in mammalian evolution (Shultz and Dunbar, 2010). Changes in brain cellular composition are attracting increasing interest (Clark et al., 2001). Reliable and comparable data for neuron and non-neuron numbers have been accumulating for several species over the last few years. This has been thanks to the development of the isotropic fractionation method (Herculano-Houzel and Lent, 2005). This method allows quantification of neurons and non-neuron cells from sections or whole brains in a reliable accurate and replicable manner with less than 10% between individual variations making data derived from this method ideal for comparative studies (Herculano-Houzel and Lent, 2005).

How different parameters have evolved and how they relate to one another has been a matter of intense study (Jerison, 2012). Larger brains tend to have higher numbers of neurons although less densely packed (Haug, 1987). The number and size of neurons, as well as the number of glial cells – an heterogeneous group of nervous system cell types other than neurons and epithelial cells– are fundamental features driving brain size increases and cognitive ability (Herculano-Houzel et al., 2014). Specifically, while neurons are generally considered to play a dominant role in determining brain function, glial cells are increasingly recognised as playing essential roles for the nervous system (Chotard and Salecker, 2004, Jessen, 2004,

Striedter, 2005, Haydon, 2001). The number of glial cells -particularly astrocytes- relative to neuron number has been proposed to be associated with increased brain complexity as the increase in density and complexity of the synaptic networks requires higher levels of astrocyte-mediated local modulation and control (Nedergaard et al., 2003). Glia number increases are closely associated with neuron numbers (Haug, 1987). Glia to neuron (glia/neuron) ratio has also been associated with decreased absolute neuronal density and increased neuronal cell body size (Herculano-Houzel, 2014). Higher glia to neuron ratio are linked to larger brain size across mammalian species (Hawkins and Olszewski, 1957, Tower and Young, 1973, Sherwood et al., 2006, Jehee and Murre, 2008) although more recent studies using better cell composition quantification have shown that glia to neuron ratio does not vary as a single function of brain size across species, but instead the relationship is clade-specific (Herculano-Houzel, 2014, Lewitus et al., 2012).

A number of studies have examined genomic signatures related to the evolution of brain complexity-related features (Castillo-Morales et al., 2014, Montgomery et al., 2011, Florio et al., 2015, Keeney et al., 2015, Gutierrez et al., 2011, Castillo et al., 2013, Castillo-Morales et al., 2016). Adaptive selection acting on two microcephaly-associated genes --*ASPM* (*MCPH5*) and *CDK5RAP2* (*MCPH3*)-- across anthropoid primates (Montgomery et al., 2011) and accelerated evolution of *MCPH1* in hominid lineages (Evans et al., 2004), suggests a role for these genes in driving the increase in brain size observed in hominid lineages. The presence of an extra copy of ARHGAP11 gene in the human genome (ARHGAP11B) not present in the mouse has been associated with basal progenitor proliferation and neocortex expansion (Florio et al., 2015). Proliferation of the DUF1220 protein domains in the primate lineage correlates with both brain size and cortical neuron number (Keeney et al., 2015). These studies suggest a role for duplication events on the cellular composition of the brain. Genome-wide comparative studies on mammalian species found an association between gene family size (GFS)

and encephalization, further supporting a role of gene duplication events in brain evolution and neocortex expansion (Castillo-Morales et al., 2014, Castillo-Morales et al., 2016).

However, genome wide determinants of brain cell composition have yet to be identified. It is not known whether any brain cell composition parameters are associated with changes in GFS. Whether previous associations between GFS and encephalization are actually explained by associations between brain cell composition parameters and GFS is also unknown. If genomic features are under selective pressure associated with cellular composition in the brain rather than on encephalization as has been proposed, then we expect that genomic features previously found to be associated with encephalization to be better explained by estimates of brain cell composition.

Here, using a comparative genomics approach we assess if changes in gene family size are associated to neuron number and density, as well as glia to neuron ratio in 19 mammalian species with fully sequenced genomes and data for brain cellular composition parameters. We then explore the extent to which resulting associations explain past findings linking GFS with encephalization. As variations in GFS and cell composition parameters can both be partly dependent on phylogenetic relatedness with more related species having higher similarity in both GFS and cell composition we apply a phylogenetic correction. We then characterize the functional annotations of gene families associated with each brain cellular composition parameter. Finally, we analyse whether there is support for the notion that gene family expansion precedes encephalization increases using phylogenetic path analyses. The results provide the first insights into the molecular basis of brain cellular composition and increase our understanding of the genomic signatures of the evolution of the mammalian brain.

Hypotheses

1. If gene family size evolution has contributed to the evolution of brain cell composition we expect to find associations between these two traits beyond chance expectations. These associations should not be explained by phylogenetic relatedness.
2. If gene family size evolution is related to brain cell composition, then we expect to find significant enrichments of functional categories beyond chance expectations.
3. If gene family size evolution underlies variations in brain cell composition then we expect to observe enrichment of functional categories related with developmental functions and specifically brain related development.
4. If gene family size evolution is related to brain cell composition then we expect any associations found not to be explained by the known association between gene family size variation and encephalization.
5. If associated gene families are related to brain development then we expect to find clusters of genes which have higher levels of expression in early development vs. postnatal stages.

Material and Methods

Cellular composition of the brain and encephalization data

Data for brain mass, body mass and cellular composition of brain tissue in 19 mammalian species, with good quality sequenced genomes (see next section) were compiled from available literature (Table 1). Only instances were cell composition data was available for the exact same species as the available genome were included in this study. Certain species for which a sequenced genome is available such as *Macropus parma*, *Aotus nancymaae*, *Saimiri boliviensis boliviensis* and *Pongo abelii* were not included as it is not clear if available cell composition data corresponded to the same sequenced species after a re-evaluation of species and subspecies

status on the Ensembl database. Data for *Pongo troglodites* and for *Gorilla gorilla* were also excluded from the analyses as reported cell counts for whole brain were not measured directly but instead extrapolated from cell counts in a subset of brain regions.

Non-neuronal cell numbers are assumed to include mostly glial cells and were used as proxy of the number of glial cell -endothelial cells have been shown to account for only a small proportion of non-neuronal cells as vascular volume is a small fraction of total brain volume (2.43-3.02%) (Herculano-Houzel, 2014, Lauwers et al., 2008). Non-neuronal cell numbers will be subsequently referred to as glia cell numbers. Glia to neuron ratio was estimated by dividing glia number by neuron number (Herculano-Houzel et al., 2007). Neuron density was calculated by dividing neuron number by estimates of whole brain mass (Herculano-Houzel et al., 2007). Encephalization index, defined as brain mass corrected by the allometric effect of body size by implementing residuals of a log–log least-squares linear regression, as calculated by Gonzalez-Lagos et al. (2010). Neuron number was log transformed to achieve a normal distribution of this parameter; all analyses presented with this variable refer to the log transformed value of neuron number.

Covariance among the four phenotypes was tested by calculating a Pearson correlation coefficients matrix (Table 2). To account for phylogenetic relatedness influencing the degree of association between different parameters, phylogenetic independent contrasts (PIC) (Felsenstein, 1985) correction was implemented using the package "ape" (Paradis and Schliep, 2019) in R (CoreTeam, 2013). Phylogenetic independent contrasts is a widely used method for assessing relatedness between variables while accounting for shared phylogenetic paths (Felsenstein, 1985).

An additional extended dataset of 56 species (Supplementary figure 1) with good quality annotated genomes but with only brain mass and body mass estimates for each species was also compiled (Supplementary table 1).

Gene family annotations

Gene family annotations were downloaded for 78 species of mammals from Ensembl-version 95 (Cunningham et al., 2018). A number of species were found to have an unusually low number of protein coding annotated genes, which are likely to result from low sequencing coverage. To test if this observation resulted from low quality genomes rather than actual variations in gene number between species we identified a set of “single copy core” genes defined as those genes present in at least 90% of species and always as a single copy. Missing a high number of these single copy core genes is more likely to be explained by low quality genome sequencing and or annotation rather than actual variations in gene number. Consistent with this, we found that those genomes with under 16000 annotated protein-coding genes are also missing a disproportionate number of the single copy core genes (Supplementary figure 2). Based on this, species with fewer than 16000 annotated protein-coding genes were not included in any analyses presented here. This removed most of the association between the single copy core gene set count and the total number of annotated genes. No further correction for total gene number was carried out in line with most studies of gene family evolution. This is because most gene families have one or two genes in each species and applying a normalisation factor would create an artificial variance in gene number between species for most gene families. The platypus was found to be an outlier in this distribution with a markedly low number of single copy core genes but having an overall number of protein coding genes typical of a mammal. This is likely to be explained by the platypus, having a different gene set or highly divergent gene sequences rather than low sequencing coverage. To avoid creating biases given its’ divergence in gene sets present, this species was removed from further analyses.

Gene families in the 19 species with available cell composition and a sequenced genome were required to have at least one gene in a minimum of 80% of species (n > 16) to be included in analyses to remove newly evolved families only present in a specific lineage. Only gene families with at least two genes in at least one species and a minimum difference of two between the maximum and the minimum number of genes across species were considered. This removed gene families with no variation in gene number across species, single copy genes only varying in their presence/absence and families where all variation is explained by a single deletion or duplication event. After applying these filters, 3,854 families were included in the study.

Phylogenetic regressions of gene family size and cell composition parameters

We used generalised least square regressions to assess the strength of associations between gene family size and phenotype variations across species. To rule out associations between phenotypic parameters and variations in GFS being explained by shared ancestry, a phylogenetically generalised least squares regression (PGLS) (Grafen, 1989, Grafen, 1992) was used. PGLS is a widely used method for assessing the association between variables in a set of species and involves constructing the phylogenetic variance-covariance matrix taking into account the phylogenetic tree which is then used to perform a generalized least squares linear regression. Phylogenetically corrected regressions were performed using the “nlme” R package (Pinheiro et al., 2018) assuming a Brownian motion model of evolution and a maximum likelihood method.

Some families were found to be associated to more than one phenotype. To account for this, functional enrichments in sets of gene families were reassessed after including all phenotypes tested in a single PGLS model. For highly co-varying phenotypes with a high number of overlapping gene

families, phylogenetic sequential regressions were used (Dale et al., 2015, Graham, 2003, Dormann et al., 2013a). For this, a PGLS model was calculated where the focal phenotype was introduced to the model alongside the residuals of the second phenotype regressed against the focal phenotype (Graham, 2003, Dormann et al., 2013a).

Effect size

Cohen's *r* effect sizes (Cohen, 1988) were calculated by computing correlation r from the *t* statistic of the PGLS model summary using the following formula (Rosenthal et al., 2000):

$$r = \frac{t}{\sqrt{t^2 + df}}$$

Where total degrees of freedom are calculated as the number of degrees of freedom in the model -total number of variables (one to four) plus the intercept minus one- subtracted from the total degrees of freedom -sample size (number of species in the test) minus one. Focusing on effect sizes, instead of on *p* values reduces the probability of type two errors, where an alternative hypothesis would be wrongly rejected, particularly when sample sizes are low (Nakagawa, 2004) as is the case in this study (n = 19) .

Associations with encephalization for gene families with large effect sizes in the set of 19 species were compared to associations for the same gene families using a larger sample of 56 species for which sequenced genome and encephalization quotient data was available. Significance of the associations for individual gene families were established for this larger set of species with Benjamini-Hochberg correction for multiple testing (Benjamini and Hochberg, 1995). The Benjamini-Hochberg correction for multiple testing is one of the most widely used FDR methods. It involves ordering the multiple tests carried out according to their *p* value from the smallest value to the largest. A moving threshold of decreasing stringency is then applied to

each test by dividing the alpha value threshold by the number of rows below. For example, setting the alpha value at 0.05, if three tests are preformed, then the significance threshold is 0.05/3 for the test with the smallest *p* value, 0.05/2 for the second test and 0.05 for the third test with the largest *p* value.

Power analysis

Power analysis was carried out to calculate the minimum number of species required to have to achieve the recommended statistical power of 0.8 (Cohen, 1988) to assess significance of large effect sizes (r > 0.5) (Cohen, 1988) when testing associations for 3,854 gene families.

For this we used the following formula from (Cummings and Hulley, 1988b) as implemented in (http://www.sample-size.net/correlation-sample-size/):

$$N = \left[\frac{Z_\alpha + Z_\beta}{C}\right]^2 + 3$$

Where Z_α is the standard normal deviate for the significance threshold 0.05 divided by the number of tests carried out; Z_β is the normal deviate of the accepted level of type two errors (0.2 for a statistical power of 0.8) and C is calculated as follows:

$$C = 0.5 * \ln\left[\frac{1+r}{1-r}\right]$$

Where *r* refers to the size of association for which significance should be reliably established.

Effective number of tests

The effective number of tests takes into account the non-independence and collinearity between individual data points in a sample to calculate the 'true'

sample size. This is calculated from the eigenvalues from a correlation matrix of gene family size profiles (Li and Ji, 2005).

Effect size distribution randomisation test

To test if the number of gene families associated with each phenotype after applying these thresholds is higher than random expectations, the number of gene families meeting this threshold was compared against the number reaching this threshold in 1,000 randomised data sets. In each randomisation, measurements for each phenotype were randomly re-assigned to species name, keeping gene family size data for each species unchanged. In a one tail test, if fewer than 50 of the randomised distributions had a larger number of families with a large effect size than the real distribution of *r*s, then the test was deemed to be significant (alpha value of 0.05 in a one-tail test).

Gene ontology term enrichment analysis

Gene ontology (GO) functional terms annotations for each gene for each species were obtained from the Gene Ontology Consortium database (www.geneontology.org). GO terms were linked to a family whenever that term was assigned to any gene in the family in any of the 78 sequenced mammalian species available in Ensembl version 95 (Cunningham et al., 2018). GO terms associated with fewer than 50 associated gene families were pooled together into a single category labelled “small GO” as overrepresentation of categories associated with very few genes would be difficult to assess and would unnecessarily reduce statistical power (Castillo-Morales et al., 2014). Unlike the case in other studies, families not associated with any functional GO term were included in the analyses under an “unknown GO” term. Enrichment of GO categories among the set of gene families associated to each of the phenotypes of interest, was carried out by measuring the proportion of families assigned to each GO term within the analysed set of gene families and comparing it with the proportion of gene

families associated to each GO term in 1,000 equally-sized samples of randomly chosen gene families from the background set. The mean and standard deviation of GO term representation as measured in each of these 1000 random samples were taken to determine the corresponding *p*-values for each GO term using *Z*-score with the formula described in the above methods section and Benjamini-Hochberg correction for multiple testing as implemented in (Castillo-Morales et al., 2014).

Co-expression and temporal gene expression analysis

To characterize and identify gene sub-groups of functionally related genes, co-expression gene networks were constructed from available human brain gene expression data. Gene expression profiles for data for 6777 genes, corresponding to the 3854 gene families were examined. These expression profiles correspond to 524 samples corresponding to 26 brain structures across human development (18 days post-conception to 40 years) from a total of 42 male and female donors were downloaded from the BrainSpan database (http://www. brainspan.org/) (Miller et al., 2014). Individuals with less than five brain structures were not included in the analysis. For each gene member of families associated with each phenotype, reads were transformed to pseudo-counts (which represent expected count number should all libraries were equal) using the EdgeR package (Robinson et al., 2010). For this, the function 'calcNormFactors' was used to calculate normalization factors in order to scale the raw library sizes. The 'estimateCommonDisp' function was then used to estimate a common dispersion value across all genes to obtain pseudo-counts.

A gene pairwise co-expression matrix was constructed by calculating Pearson correlation coefficients for expression profiles for each gene against all others. Weighted gene co-expression network analysis was carried out using WGCNA R package (Langfelder and Horvath, 2008) to identify clusters of co-expressed genes. In order to detect modules of highly co-expressed

genes within every list, unsupervised hierarchical clustering was performed following the method described by (Zhang and Horvath, 2005). In each case, soft power values were defined as the first value when reaching a scale free topology fitting index and passing a $R^2 > 0.8$ threshold using the function 'pickSoftThreshold'(Zhang and Horvath, 2005). A threshold of a minimum of 20 genes per module was used. For genes in every resulting cluster, mean brain expression for at each development time point was then calculated. One-tailed *t*-test per cluster was carried out to compare a cluster's mean activity in prenatal and postnatal stages. Benjamini-Hochberg correction was used to correct for multiple testing.

Phylogenetic path analysis

In order to determine the causal relationship between encephalization and size of associated gene families, phylogenetic path analysis (Gonzalez-Voyer and Von Hardenberg, 2014) was carried out using the R package "phylopath" (van der Bijl, 2018) under the default model (Pagel's lambda). The "phylo_path" function, which compares causal hypotheses and corrects by the phylogenetic relationships among species was used. The analysis was only carried out for encephalization in the larger set of 56 mammalian species and using body mass as a control variable as the smaller set of 19 species for which cell composition was available is too small a dataset for this analysis. To this end, the number of genes in all families associated with encephalization were added to obtain a single value per species. Then two hypotheses were tested: H_0 that changes in total number of genes among encephalization associated gene families drive changes in body mass. H_1 that changes in total number of genes among encephalization associated gene families drives changes in encephalization.

Results

We started by evaluating the relationship between brain cell composition parameters and gene family size (GFS). To this end, GFS and parameters of brain cell composition were compiled for each family in 19 mammalian species for which cell composition data and fully sequenced genomes were available (Table 1). Using phylogenetically corrected correlation analysis we found that more encephalized species have higher neuron numbers ($r = 0.84$, $p < 0.001$) as previously reported (Herculano-Houzel et al., 2007). Significant associations between glia to neuron ratio and neuron density ($r = 0.57$, $p = 0.008$), between neuron number and neuron density ($r = 0.47$, $p = 0.037$) and neuron and glia numbers were found to be strongly correlated ($r = 0.99$, $p < 0.001$; Table 2). Given the high collinearity between neuron and glia number, only results for log neuron number (referred as “neuron number”) are reported here, to avoid unreliable results. The variations in glia number still contribute to the glia to neuron ratio.

Next, the association between gene family size with encephalization, neuron number, glia to neuron ratio and neuron density was assessed through regression models. To exclude the effect of shared evolutionary paths among species influencing the association between gene family size and phenotype variations, phylogenetic generalised least squares (PGLS) analysis was used. Using a large effect size threshold ($r > 0.5$), a total of 291, 51 and four gene families were found to be positively associated to neuron number, glia to neuron ratio and neuron density, respectively. Encephalization was found to be associated with 447 gene families (Figure 1).

As our species sample size is small, we do not have the statistical power to evaluate significance of associations for 3854 individual gene families. Indeed, based on power analysis calculations, data for at least 93 species would be required to be able to evaluate significance of even large effect

sizes (r > 0.5) (Cohen, 1988) and based on the effective number of tests (Li and Ji, 2005) which estimated an effective number of tests on 3554, data for at least 92 species would be needed. As expected, no significant associations were found between GFS and the four parameters tested in the set of 19 species after correcting for multiple testing. Taking advantage of an expanded set of species (n = 56) for which encephalization (but not cell composition) data was available, we found that the majority of families classed as associated with encephalization (79%) in the smaller data set based on effect size, were significantly associated with encephalization after Benjamini-Hochberg correction in the set of 56 species. Thus, a threshold based on a large effect size (r > 0.5) adequately selects sets of gene families enriched in significant associations when testing them in a larger sample size.

Furthermore, when examining the distribution of correlation coefficients, the number of gene families with large effect size associations is significantly higher than random expectations for encephalization (for the smaller and larger datasets: n = 19, p = 0.033; n = 56, p = 0.006) and neuron number (n = 19, p = 0.027). No such excess was found, however, for glia to neuron ratio and neuron density (p > 0.05). Importantly, the excess of associations with a regression coefficient higher than 0.5 between GFS for neuron number or encephalization, are not explained by an overall association between the total number of annotated protein coding genes per genome. Total number of protein coding genes per genome was not significantly associated with encephalization, neuron number, glia to neuron ratio and neuron density (PGLS: n = 19, p > 0.05, r < 0.5).

Functional characterisation of associated gene families for each phenotype showed that encephalization and neuron number are both significantly enriched in immune related functions with encephalization also enriched in synaptic plasticity (Figure 2a). Translation associated gene families were

enriched among families associated with neuron density (although this is based on only four gene families associated with this phenotype). Neuron ratio was associated with steroid metabolism. Enrichment in immune related functions has been previously reported among gene families associated with encephalization (Castillo-Morales et al., 2014).

We further investigated if the association between GFS related to immune system functions with encephalization and neuron number could be explained by covariance between the two phenotypes (r = 0.86, p < 0.001; Table 2). Indeed, there were a total of 267 shared gene families out of 291 and 447 of families associated to neuron number and encephalization, respectively (Figure 1). Given the high collinearity between neuron number and encephalization, we performed a sequential regression (Graham, 2003, Dormann et al., 2013a) to reassess sets of gene families associated with each phenotype, first using encephalization and then neuron number as focal variables. A total of 482 and 398 gene families were found to be significantly associated with encephalization and neuron number respectively after removing the effect of the other variable. Both sets of gene families associated with encephalization and neuron number were still found to be associated with immune system related functions (Figure 2b).

In addition to the strong correlation between encephalization and neuron number, significant associations were also observed between glia to neuron ratio and neuron density (PIC: r = 0.585, p = 0.011) and a weaker association between neuron density and neuron number (r = 0.522, p = 0.026; Table 2). To account for shared explained variance, all four phenotypes were included into a single PGLS model. Given the strong collinearity between neuron number and encephalization, residuals of neuron number regressed against encephalization were included into the model instead of neuron number. Neuron number was included as residuals as the largest number of gene families was found to be associated with

encephalization in the reciprocal sequential regression analysis including the two variables only. In this model, we found 1,147 gene families associated with encephalization, 534 with neuron number residuals, 61 associated with neuron density and 35 with glia to neuron ratio.

Functional enrichments among families associated with each phenotype were reassessed after all four phenotypes in this study were included into a single PGLS model. Encephalization and neuron number residuals-associated gene families were found to be independently enriched in immune system-related functions (Figure 2c). Interestingly, gene families associated with encephalization after removing the effect of shared variance with other phenotypes were significantly enriched in brain function and brain development-related processes. Gene families associated to neuron number residuals were enriched with families annotated to synaptic plasticity, cell cycle and cell to cell signalling in addition to immune system functions. Significant enrichments in translation and cell migration for families associated to glia to neuron ratio were found. Enrichment in cell projection and neuron development functional categories were observed among families associated with neuron density. These results further support the existence of significant functional enrichments in immune system function among gene families associated to both neuron number and encephalization which are not explained by co-variance among the two variables or with other cell composition phenotypes or phylogenetic relatedness.

To further characterise gene families associated with encephalization and cell composition parameters, gene expression patterns in the human brain across age groups from 18 days post conception to 40 years of age were examined. This is important as genes with higher expression in the brain during prenatal stages are likely to be playing roles during development of the brain which could influence the parameters here studied. For this, a weighted co-expression network was constructed from expression patterns

of all gene members from families associated with each phenotype (2,907 genes associated with encephalization, 1,195 with neuron number, 60 with glia / neuron ratio and 107 with neuron density). Mean expression across time was then calculated for each resulting cluster of co-expressed genes and each were functionally characterised. A total of 25 clusters were identified for encephalization associated genes, 12 for neuron number and one cluster for neuron density and glia to neuron ratio. Several clusters were found to have a marked upregulation in prenatal stages compared to postnatal ones (Figure 3).

Finally, to assess the likelihood that gene family expansion has contributed to the evolution of brain size we used a phylogenetic path analysis test. We found that increases in gene number in the set of encephalization-associated gene families drove subsequent increases in encephalization and not on the control variable body mass (p = 0.721 and p = 0, respectively; significance indicates rejection; Figure 4). Given the smaller sample size of 19 species for which data on cell composition is available, path analysis was not performed in gene family sets associated with these phenotypes.

Discussion

Cross-species variations in gene family size can reveal changes in the relative evolutionary relevance of distinct molecular functions. By using a comparative approach, we can identify key genomic determinants and molecular functions associated to the evolution of larger and more complex brains such as that of humans; something that would otherwise be impossible using conventional individual animal model-based approaches.

This study examined the association between cellular composition (neuron number, glia to neuron ration and neuron density) in the mammalian brain and changes in gene family size. In line with hypothesis one, we found an excess of gene families associated with encephalization and total neuronal count compared to chance expectations and distinct sets of functional categories associated with these groups of families. These associations were not explained by phylogenetic relatedness. Furthermore, an overall association between these two phenotypes and total annotated gene numbers per species does not explain the excess of expanding gene families in association with encephalization and neuron number.

Gene families expanding in line with encephalization and neuron number were found to be associated with immune system functions as well as brain development related functions in accordance with expectations from hypotheses 2 and 3. Expansion of gene families associated with the immune system in line with encephalization has been previously shown in a smaller set of species (Castillo-Morales et al., 2014, Castillo-Morales et al., 2016). If the association between encephalization and gene family variation is a by-product of selection favouring gene family expansion with neuron number, then we expect previously reported associations between GFS and encephalization to be accounted for by associations between GFS and neuron number. Crucially, sequential regressions assessing GFS association to encephalization and neuron number revealed that both neuron number and encephalization are independently associated with sets of gene families enriched in immune system genes. General linear models including both parameters in combination with other cell composition parameters (neuron density and glia to neuron ratio) further confirm that gene families associated with neuron number and encephalization are independently associated with immune related functions. These results were in agreement with hypothesis four.

Many regulatory and signalling components of the immune system have been shown to play important roles in the nervous system development (Monzón-Sandoval et al., 2015) which may account for the overrepresentation of immune system gene functions. This may result from co-evolution of both the nervous and the immune system supported by common molecular pathways. It is also possible that selective pressures for transcript diversification in both systems have favoured co-option of molecular pathways from one system to the other. Future analyses examining phenotypic parameters of immune system evolution could shed light into this. The expansion of immune related gene families in line with encephalization and neuron number might also be explained by the co-evolution of brain size and placenta invasiveness as larger brain to body ratios have been associated with more invasive placentas (Elliot and Crespi, 2008) although strong associations between brain size and placenta structural features are not observed (Montiel et al., 2013). Thus, it is possible that the expansion of immune system genes in line with larger brains might be an indirect result of the immune challenge that more invasive placentas associated with larger and more encephalized species represent for the mother (Bainbridge, 2000).

Significant enrichments among neuron number and encephalization-related families were observed for several other categories shown to be key for brain development including cell-cell signalling and regulation of synaptic plasticity (Rodrigues et al., 2018, Seiradake et al., 2014, Budnik and Salinas, 2011, Mateos-Aparicio and Rodríguez-Moreno, 2019). Encephalization and neuron density-associated gene families were enriched in negative regulation of neuron differentiation annotations another key neural specific function (Mariani et al., 2012, Sun et al., 2017). Neuron density showed further over-representation of gene families associated with neuron development, microtubule cytoskeleton organization, cilium assembly, with the latter two known to be crucial for neural connectivity establishment during development (Menon and Gupton, 2016, Park et al., 2019). Finally, positive regulation of

cell migration and regulation of cell proliferation functions are over-represented among gene families associated to encephalization and glia to neuron ratio. These processes are part of the fundamental mechanisms underlying the development of functional neural circuits in the brain (Marín et al., 2010). Cell migration, for instance, allows the appropriate positioning of cell types from different origins and enables the generation of brain circuitries (Marín et al., 2010). Neuronal migration is important for the establishment and maintenance of functional neuronal connectivity in the developing and adult brain (Evsyukova et al., 2013). Moreover, cell proliferation is a process associated to the maintenance of glial cell numbers in the adult brain as well as neurons during development. The balanced regulation of this mechanism and cell dead is crucial for the correct number of specific cell types and final size of the brain (Joseph and Hermanson, 2010). Furthermore, it has been proposed that in eutherian evolution, the mechanisms of change that led to diversity and conservation in brain scaling involved clade and brain region-specific modifications in the pathways that regulate cell proliferation and cell death (Herculano-Houzel et al., 2014).

Examining individual gene families associated with encephalization we can highlight the BCL2, SHANK and MDB gene families as examples. BCL2 proteins are differentially expressed during the different stages of brain development (Opferman and Kothari, 2017) and are involved in the regulation of neuronal programmed cell death, are required for cortical neurogenesis and the survival of neurons after DNA damage (Arbour et al., 2008). Furthermore, their role in neurodegenerative processes including Alzheimer's, Parkinson's and Huntington's diseases and Amyotrophic Lateral Sclerosis, has been widely studied (Akhtar et al., 2004, Sassone et al., 2013). The gene family SHANK encoding post-synaptic scaffold proteins has been associated with a variety of nervous system and brain development functions (cell differentiation, axon guidance, nervous system and brain development, learning, memory, regulation of synaptic plasticity, and synapse assembly). Interestingly, mutations on SHANK genes are involved

in synaptic dysfunction (Wang et al., 2011) and the development of neurodevelopmental and neuropsychiatric disorders (Leblond et al., 2012, Sala et al., 2015). The methyl-CpG binding domain (MBD) family, associated with encephalization and neuron number, is involved in transcriptional repression and has been shown to play important roles in neuronal development (Jung et al., 2003, Knock et al., 2015, Zhao et al., 2003), and dysfunction of genes in this family has been associated with psychiatric disorders (Gigek et al., 2016). It would be interesting to examine whether expanding gene families in line with brain evolution are disproportionately associated with psychiatric disorders and/or neurodegeneration.

Further supporting the role of genes associated with brain development, and in agreement with expectations under hypothesis five, after constructing co-expression derived gene clusters, we found several of them associated with neuron number and encephalization being overexpressed in prenatal stages compared to postnatal stages including adults. One such cluster was also found for glia to neuron ratio and neuron density. Future analyses examining the functional nature of clusters of genes within the set of phenotype associated sets of gene families. Such analyses could also help examine the likely function in different developmental states of candidate genes.

It is also important to note that the present study aimed to answer whether evolution of relative brain size and cellular composition are associated with variations in gene family size beyond chance expectations; current sample sizes do not allow confirmation of associations of individual gene families. It is worth noting that where sample sizes allowed, then phylopath analysis results are consistent with a causal link between changes in gene number increases in encephalization associated gene families but not for the control variable body mass. As additional data on cell composition accumulates in the future, it will be possible to test causal associations in other phenotypes.

It is possible that the observed excess of expanding gene families in line with encephalization and neuron number results from selective pressures favouring transcript diversification to overcome higher selective constraints and allow greater optimisation of more narrowly expressed transcripts. As the mammalian brain is a highly diverse tissue composed of hundreds of cell types and complex connectivity, we can expect brain-expressed genes to be under higher selective constraints (Kuma et al., 1995, Wang et al., 2006). Thus, any new allele, even if beneficial in a particular cellular context is likely to be deleterious in another where the gene is also expressed (Kuma et al., 1995, Wang et al., 2006). Indeed, several studies have found that brain expressed genes tend to be slow evolving (Kuma et al., 1995, Duret and Mouchiroud, 2000) with this effect being more pronounced in highly encephalized humans compared with less so primates (Wang et al., 2006). Retention of duplicate genes allows optimising each gene to the conditions in a narrower area or cell types compared to the single copy gene. As more encephalized species have a larger neocortex and a greater diversity of structural areas (Kaas, 1993, Roth and Dicke, 2005), selective constraints would be higher on brain expressed genes in highly encephalized species favouring retention of duplicate genes compared to species with lower relative brain size. The high levels of alternative splicing in the brain (Yeo et al., 2004) offers further support to the transcript diversification pressure hypothesis. In this regard, a combined analysis of alternative splicing and gene duplication evolution in line with brain evolution would be an interesting area to explore.

In summary, we present evidence of an association between variations in gene family size and differences in key brain cellular composition parameters known to contribute to nervous system functions. We found an excess of associations between the number of genes per species with encephalization on the one hand and neuron number on the other after correction for phylogenetic relatedness compared to chance expectations. These sets of gene families were independently enriched in immune system-related

biological functions. Enrichment of gene families associated with various aspects of brain development were found in expanding gene families in line with encephalization as well as neuron number, glia to neuron ratio and neuron density. Together, our results represent the first genomic screening for gene families associated with evolutionary variations in cellular composition of the brain in mammals.

Tables

TABLE 1. Data for brain composition features for mammalian species with fully sequenced reference genome available.

Species	*Ei*	Neuron number, x 10^6	Neuron density, N/g	Glia/neuron ratio
Bushbaby	-2.12	936[(1)]	92.22	0.71
Cat	-1.82	1215.21[(2)]	34.86	1.42
Crab-eating macaque	-1.25	3440[(3)]	74.5	0.92
Dog	-2.17	2252.69[(2)]	26.05	2.31
Elephant	-1.35	257043.4734[(4)]	55.65	0.84
Ferret	-2.71	404.43[(2)]	74.49	1.17
Golden Hamster	-3.07	84.22[(5)]	87.27	0.84
Guinea Pig	-2.76	233.56[(5)]	63.88	0.98
Human	0.15	86060[(6)]	57.07	0.98
Macaque	-1.11	6376.16[(1)]	73	1.12
Marmoset	-1.71	635.8[(1)]	81.72	0.93
Mouse	-2.67	67.87[(5)]	168.83	0.5
Mouse Lemur	-2.32	254.71[(3)]	141.58	0.55
Naked mole-rat male	-3.26	26.88[(5)]	68.57	0.9
Olive baboon	-0.98	10950[(3)]	72.43	0.84
Pig	-1.95	2220[(7)]	37.9	2.08
Rabbit	-2.42	494.2[(5)]	54.12	1.28
Rat	-2.8	188.87[(5)]	109.55	0.65
Tasmanian devil	-2.77	693.69[(8)]	70.77	1.09

Ei = Encephalization index; N = Neuron number. [(1)] Herculano-Houzel *et al*. (2007) [(2)] Jardim *et al*. (2017). [(3)] Gabi *et al*. (2010). [(4)] Herculano *et al*. (2015). [(5)] Herculano-Houzel *et al*. (2011). [(6)] Azevedo *et al*. (2009). [(7)] Kazu *et al*. (2014). [(8)] Dos Santos *et al*. (2017). Species scientific names are in Supplementary Table 1.

TABLE 2. Relationship between cell composition and encephalization in mammals. Autocorrelation matrix showing the correlation coefficients between different brain features. The corresponding *p*-values are shown inside parentheses.

	Encephalization	**Glia number**	**Neuron number**	**Glia/neuron ratio**
Glia number	0.842 (0.000)			
Neuron number	0.858 (0.000)	0.988 (0.000)		
Glia/neuron ratio	0.279 (0.263)	0.475 (0.046)	0.345 (0.161)	
Neuron density	0.357 (0.146)	0.603 (0.008)	0.522 (0.026)	0.585 (0.011)

Figure legends

FIGURE 1. Gene family associations with encephalization and cell composition parameters. Venn diagram comparing gene families displaying a positive association between GFS and each phenotype tested: encephalization, neuron number neuron density and glia to neuron ratio.

FIGURE 2. Functional annotation overrepresentation among gene families associated with each phenotype. Heatmaps show gene ontology (GO) biological process category enrichments among members of gene families associated with different cell composition features (neuron number, neuron density and glia to neuron ratio) and encephalization. Only significantly overrepresented terms in at least one model (with a cut off threshold for significance of 0.05 and using Benjamini Hochberg FDR correction for multiple testing) are shown. Panel **a)** shows enrichments for gene families associated with each phenotype when input into de model on their own. Panel **b)** show enriched terms among gene families associated with encephalization and or neuron number in PGLS models where each variable takes its turn as focal variable with the residuals of the other. Panel **c)** shows enriched terms in among families associated with each variable tested when all are included into a single model.

FIGURE 3. Gene cluster dendrograms and temporal changes in mean gene expression of genes in selected clusters across human development. Dendrograms and co-expression modules for selected clusters derived from co-expression analyses among genes associated with each phenotype. Panels show gene expression mean activity across development for selected clusters for **a)** encephalization, **b)** neuron number and **c)** glia to neuron ratio. Neuron density as a variable was not included here as there were no clusters showing a drop in mean expression after

birth. Temporal patterns of gene expression of three representative colour modules: **d)** blue in encephalization, **e)** black in neuron number, and **f)** the turquoise in glia to neuron ratio. Birth point is indicated with a dashed line. Full analyses of gene expression for all modules is shown in supplementary material (Supplementary tables 2-5 and Supplementary figures 3-6).

FIGURE 4. Phylogenetic path analysis. a) Schema depicting the hypotheses tested. H_0 represents the influence of gene family size of families associated with encephalization ($GFS_{Encephalization}$) over mean body mass values while H_1 stands for the effect of $GFS_{Encephalization}$ over encephalization. **b)** Statistical values resulting from the phylo_path function for the hypotheses tested. **c)** Histogram showing model weights. The bars represent the *p*-value for each of the hypotheses where significance indicates rejection (exact values shown at the end of the bars). The bar for H_1 is shown in blue as is the only hypothesis with a value of delta CICc lower than two (as shown in b). Both criteria, *p*-value and delta CICc indicate the significance of a hypothesis.

Figures

FIGURE 1.

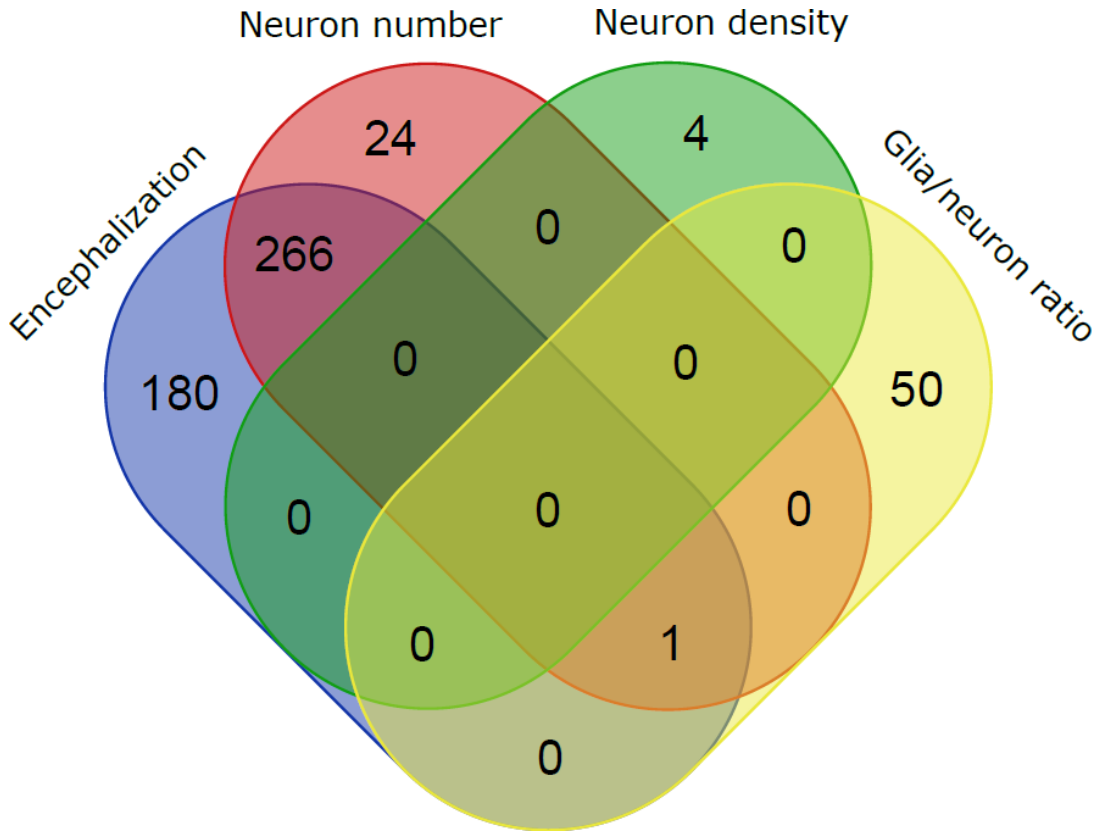


FIGURE 2.

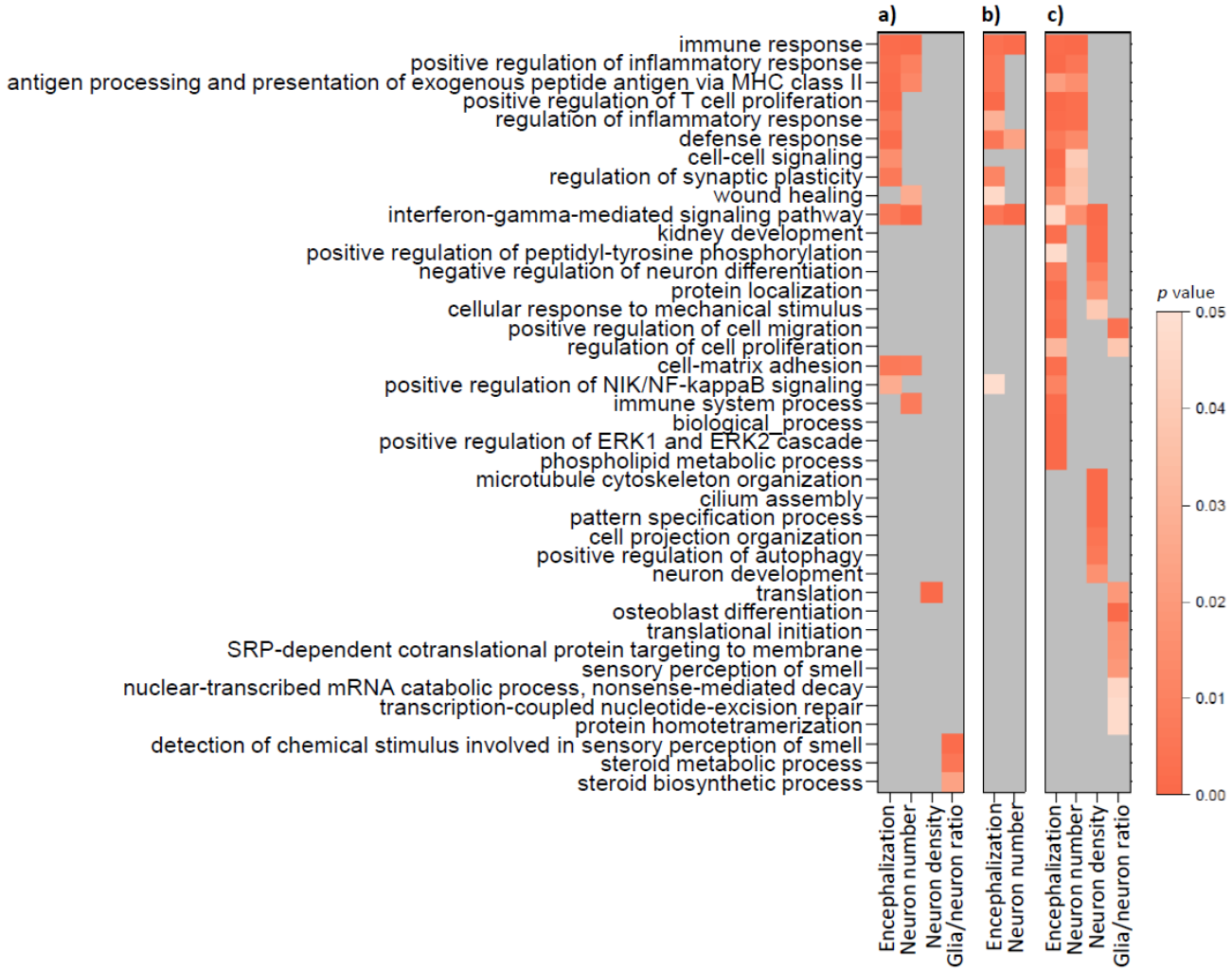

a)
b)
c)
immune response
positive regulation of inflammatory response
antigen processing and presentation of exogenous peptide antigen via MHC class II
positive regulation of T cell proliferation
regulation of inflammatory response
defense response
cell-cell signaling
regulation of synaptic plasticity
wound healing
interferon-gamma-mediated signaling pathway
kidney development
positive regulation of peptidyl-tyrosine phosphorylation
negative regulation of neuron differentiation
protein localization
cellular response to mechanical stimulus
positive regulation of cell migration
regulation of cell proliferation
cell-matrix adhesion
positive regulation of NIK/NF-kappaB signaling
immune system process
biological_process
positive regulation of ERK1 and ERK2 cascade
phospholipid metabolic process
microtubule cytoskeleton organization
cilium assembly
pattern specification process
cell projection organization
positive regulation of autophagy
neuron development
translation
osteoblast differentiation
translational initiation
SRP-dependent cotranslational protein targeting to membrane
sensory perception of smell
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
transcription-coupled nucleotide-excision repair
protein homotetramerization
detection of chemical stimulus involved in sensory perception of smell
steroid metabolic process
steroid biosynthetic process
Encephalization
Neuron number
Neuron density
Glia/neuron ratio
Encephalization
Neuron number
Encephalization
Neuron number
Neuron density
Glia/neuron ratio
p value
0.05
0.04
0.03
0.02
0.01
0.00

FIGURE 3.

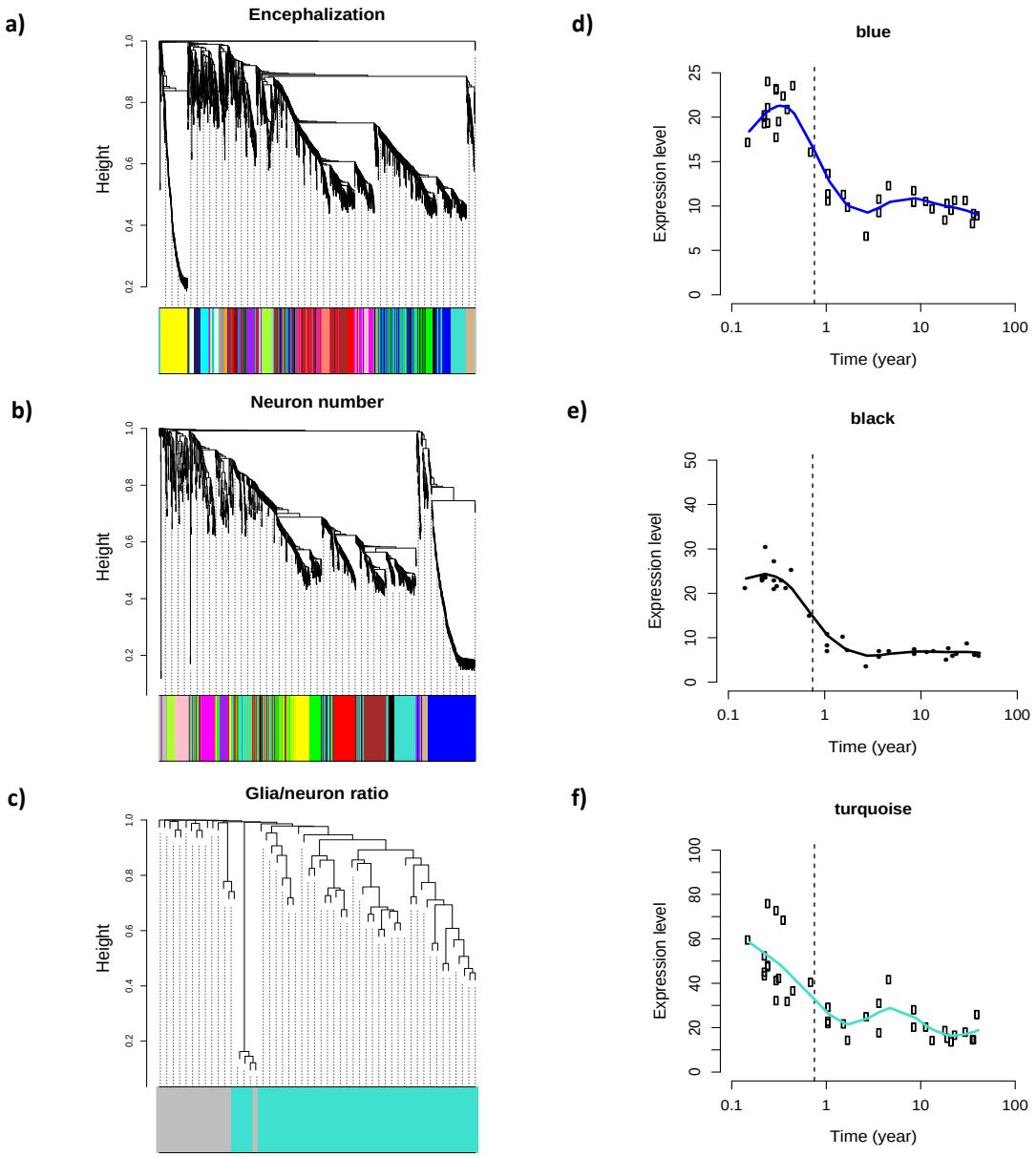
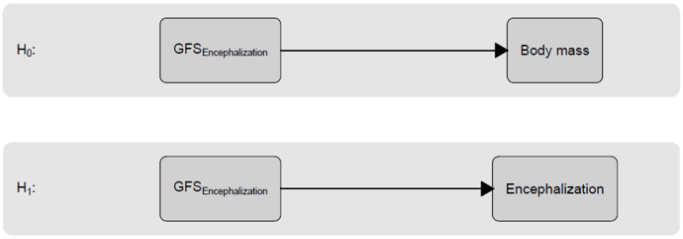

a)
Encephalization
Height
b)
Neuron number
Height
c)
Glia/neuron ratio
Height
d)
blue
Expression level
Time (year)
e)
black
Expression level
Time (year)
f)
turquoise
Expression level
Time (year)

FIGURE 4.

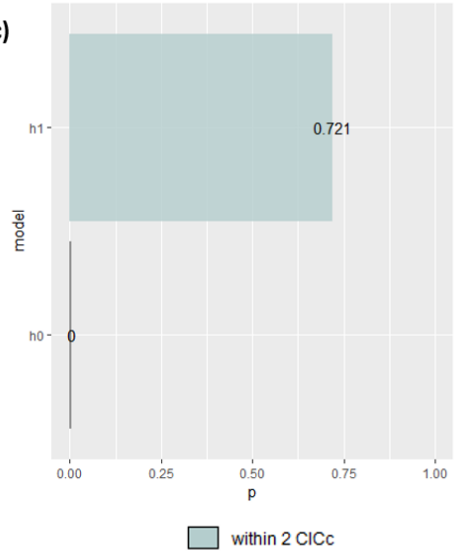
a)



b)

model	k	q	C	p	CICc	delta_CICc	l	w
h_1	2	4	2.081497	0.7207724	10.86581	0	1	0.9999999
h_0	2	4	35.31533	4.001576e−07	44.09964	33.23383	6.072466e−08	6.072466e−08

c)



Supplementary Material

TABLE S1. Data for brain and body mass for the mammalian species with Encephalization data and fully sequenced reference genome available.

Common name	Species name	Encephalization	Brain mass, g	Body mass, g
American black bear	*Ursus americanus*	-1.71	228[1]	70025[16]
Angola colobus	*Colobus angolensis palliatus*	-1.48	74.4[2]	8525[16]
Armadillo	*Dasypus novemcinctus*	-2.91	12[3]	4550[17]
Bolivian squirrel monkey	*Saimiri boliviensis boliviensis*	-1.14	24.1[2]	857.5[18]
Bonobo	*Pan paniscus*	-0.91	326.3[1]	35000[18]
Bushbaby	*Otolemur garnettii*	-2.12	10.45[4]	1074.65[19]
Capuchin	*Cebus capucinus imitator*	-0.9	72.1[1]	3267[18]
Cat	*Felis catus*	-1.82	28.4[3]	3200[20]
Chimpanzee	*Pan troglodytes*	-0.94	371.05[3]	44750[21]
Chinese hamster	*Cricetulus griseus*	-2.67	0.628[2]	31.2[2]
Cow	*Bos taurus*	-2.23	456[3]	462500[22]
Crab-eating macaque	*Macaca fascicularis*	-1.25	62.1[1]	4461.5[16]
Degu	*Octodon degus*	-2.78	2.1[1]	245[23]
Dog	*Canis familiaris*	-2.17	100.9[5]	40000[24]
Dolphin	*Tursiops truncatus*	-0.47	1679.6[6]	226700[25]
Donkey	*Equus asinus asinus*	-1.63	389.1[1]	141500[26]
Elephant	*Loxodonta africana*	-1.35	4480[3]	4150000[18]
Ferret	*Mustela putorius furo*	-2.71	7.1[3]	1472.5[27]
Gibbon	*Nomascus leucogenys*	-0.92	119.4[7]	7365[28]
Goat	*Capra hircus*	-2.26	115[1]	56786[22]

Common name	Species name	Encephalization	Brain mass, g	Body mass, g
Golden Hamster	*Mesocricetus auratus*	-3.07	0.965[6]	114.5[29]
Gorilla	*Gorilla gorilla*	-1.44	438.18[3]	126500[21]
Guinea Pig	*Cavia porcellus*	-2.76	4.28[3]	721.85[21]
Horse	*Equus caballus*	-1.75	650.03[8]	382666.5[22]
Human	*Homo sapiens*	0.15	1400[3]	65000[21]
Kangaroo rat	*Dipodomys ordii*	-1.88	1.97[7]	54.01[21]
Koala	*Phascolarctos cinereus*	-2.89	19.2[9]	9300[30]
Leopard	*Panthera pardus*	-2.4	112[1]	67500[31]
Lesser Egyptian jerboa	*Jaculus jaculus*	-2.37	1.21[2]	54.5[32]
Long tailed chinchilla	*Chinchilla lanigera*	-1.94	6.8[1]	417[16]
Macaque	*Macaca mulatta*	-1.11	97.45[10]	7249.25[16]
Marmoset	*Callithrix jacchus*	-1.71	7.24[10]	319.95[33]
Megabat	*Pteropus vampyrus*	-2.14	9.53[11]	959.5[21]
Microbat	*Myotis lucifugus*	-3.08	0.175[11]	8.055[34]
Mouse	*Mus musculus*	-2.67	0.45[3]	18.7[21]
Mouse Lemur	*Microcebus murinus*	-2.32	1.68[10]	85[18]
Naked mole-rat male	*Heterocephalus glaber*	-3.26	0.392[4]	37.75[35]
Northern American deer mouse	*Peromyscus maniculatus bairdii*	-2.31	0.6[1]	16.55[36]
Olive baboon	*Papio anubis*	-0.98	201[10]	18440.75[16]
Opossum	*Monodelphis domestica*	-2.97	0.95[7]	95[16]
Orangutan	*Pongo abelii*	-1.17	343[7]	56750[21]
Panda	*Ailuropoda melanoleuca*	-1.95	235.1[12]	107000[37]
Pig	*Sus scrofa*	-1.95	180.2[6]	70850[22]
Pig-tailed macaque	*Macaca nemestrina*	-1.18	100.7[1]	8500[18]

Common name	Species name	Encephalization	Brain mass, g	Body mass, g
Polar bear	*Ursus maritimus*	-2.23	365[(1)]	329550[(38)]
Prairie vole	*Microtus ochrogaster*	-2.74	0.7[(1)]	41.65[(39)]
Rabbit	*Oryctolagus cuniculus*	-2.42	9.14[(8)]	1383[(21)]
Rat	*Rattus norvegicus*	-2.8	2.38[(6)]	310[(21)]
Red fox	*Vulpes vulpes*	-1.78	43.4[(1)]	5800[(40)]
Sheep	*Ovis aries*	-1.78	103.93[(6)]	23000[(18)]
Sooty mangabey	*Cercocebus atys*	-1.31	86.8[(1)]	8300[(18)]
Squirrel	*Ictidomys tridecemlineatus*	-2.14	3.2[(13)]	175[(21)]
Tarsier	*Carlito syrichta*	-1.84	3.5[(14)]	125[(18)]
Tasmanian devil	*Sarcophilus harrisii*	-2.77	16.26[(15)]	5875[(21)]
Tiger	*Panthera tigris altaica*	-2.26	247[(1)]	187500[(31)]
Vervet-AGM	*Chlorocebus sabaeus*	-1.28	65.2[(2)]	5049.75[(16)]

(1) (Isler and van Schaik, 2012). (2) (Boddy et al., 2012). (3) (Garwicz et al., 2009). (4) (Herculano-Houzel et al., 2011). (5) (Tacutu et al., 2018). (6) (Sacher and Staffeldt, 1974). (7) (Morales, 2015). (8) (Wang et al., 2008). (9) (De Miguel and Henneberg, 1998). (10) (Leonard et al., 2007). (11) (Stephan et al., 1981). (12) (Gittleman, 1986). (13) (McNab and Eisenberg, 1989). (14) (Morand and Ricklefs, 2005). (15) (Weisbecker and Goswami, 2010). (16) (Research, 2016b). (17) (M. Mcdonough, 2000). (18) (Weckerly, 1998). (19) (O'Mara et al., 2012). (20) (Yamane et al., 1996). (21) (Jones et al., 2009). (22) (Mysterud, 2000). (23) (Suarez and Mpodozis, 2009). (24) (Société des Produits Nestlé S.A., 2019). (25) (Read et al., 1993). (26) (Nengomasha et al., 2016). (27) (He et al., 2002). (28) (Smith and Jungers, 1997). (29) (Gattermann et al., 2002). (30) (Nowak and Dickman, 2005). (31) (Lerner, 2013). (32) (Happold, 1975). (33) (Araújo et al., 2000). (34) (Kurta and Kunz, 1988). (35) (Pinto et al., 2010). (36) (Wolff, 1985). (37) (Charlton et al., 2009). (38) (Kingsley, 1979). (39) (Bondrup-Nielsen and Ims, 1990). (40) (Macdonald and Sillero-Zubiri, 2004).

TABLE S2. Encephalization. Clusters obtained from the weighted gene co-expression network analysis among the encephalization associated genes. *p*-values and adjusted *p*-values (Benjamini-Hochberg correction) from the *t*-test performed are shown, as well as mean prenatal gene expression and mean postnatal gene expression.

Cluster id	*Gene number*	*Prenatal expression*	*Postnatal expression*	*p-value*	*Adj. p-value*
Blue	300	20.5740815	10.2390919	4.70E-13	1.17E-11
Green	234	9.52754119	6.28156688	1.39E-09	3.33E-08
Lightgreen	48	12.8677773	10.7639164	0.00040857	0.00939708
Darkgrey	23	10.1893915	7.73689135	0.00080769	0.01776915
Tan	84	11.7195682	9.55130354	0.00220079	0.04621659
Darkred	32	11.1302124	9.29613296	0.00279093	0.0558185
Yellow	242	0.14807533	0.12352518	0.01715851	0.32601176
Orange	23	10.2611262	9.74919962	0.15186569	2.7335825
Greenyellow	87	10.7008521	10.4071174	0.26617727	4.52501354
Lightyellow	37	9.99757056	9.77451766	0.36007075	5.76113198
Lightcyan	51	10.4825113	10.345714	0.43621723	6.54325839
Grey60	50	5.91745551	5.88011487	0.44840265	6.27763703
Darkturquoise	28	33.2744484	33.0141791	0.45519666	5.91755654
Cyan	73	14.2504395	14.8562761	0.71961489	8.63537867
Red	204	9.1159483	9.59324945	0.84258475	9.26843222
Midnightblue	57	11.7860164	12.8043367	0.92787431	9.27874308
Royalblue	33	19.5051781	23.0788143	0.9787927	8.80913426
Salmon	78	10.2028022	11.4091627	0.99073315	7.92586517
Brown	285	13.2109571	15.3550281	0.99280898	6.94966283
Turquoise	321	12.3921883	15.5924776	0.9944873	5.96692382

Cluster id	Gene number	Prenatal expression	Postnatal expression	*p*-value	Adj. *p*-value
Black	192	9.02783524	10.3482841	0.99512874	4.9756437
Darkgreen	31	9.76910088	11.3303819	0.9955382	3.98215278
Magenta	143	10.9275333	13.8471085	0.99989387	2.99968162
Purple	98	11.466626	16.8582771	0.99997642	1.99995285
Pink	148	5.99167996	9.23867878	0.99999977	0.99999977

TABLE S3. Neuron number. Clusters obtained from the weighted gene co-expression network analysis among the neuron number associated genes. *p*-values and adjusted *p*-values (Benjamini-Hochberg correction) from the *t*-test performed are shown, as well as mean prenatal gene expression and mean postnatal gene expression.

Cluster id	Gene number	Prenatal expression	Postnatal expression	*p*-value	Adj. *p*-value
Black	81	23.204367	7.12328099	3.62E-13	4.35E-12
Purple	52	6.46755296	5.34152501	2.11E-05	0.00023235
Blue	184	0.12592023	0.11772007	0.02990766	0.29907655
Brown	138	7.06637838	6.72770337	0.051274	0.46146603
Yellow	125	9.58369006	8.91376644	0.06854562	0.54836495
Pink	54	8.95555139	8.80757467	0.33654389	2.3558072
Greenyellow	33	20.148571	21.7856758	0.80513396	4.83080373
Red	104	8.98851703	9.37076708	0.82005458	4.10027288
Tan	23	0.30230225	0.3321943	0.88640001	3.54560002
Turquoise	219	9.22943421	10.3333492	0.97740935	2.93222804
Green	107	7.50964803	8.77733261	0.99885426	1.99770851
Magenta	53	14.1085316	21.640762	0.99997335	0.99997335

TABLE S4. Glia / neuron ratio. Clusters obtained from the weighted gene co-expression network analysis among the glia / neuron ratio associated genes. *p*-values and adjusted *p* - values (Benjamini-Hochberg correction) from the *t*-test performed are shown, as well as mean prenatal gene expression and mean postnatal gene expression.

Cluster Id	Gene number	Prenatal expression	Postnatal expression	*p*-value	Adj *p*-value
Turquoise	45	49.4252003	21.4800604	4.55E-07	4.55E-07

TABLE S5. Neuron density. Clusters obtained from the weighted gene co-expression network analysis among the neuron density associated genes. *p*-values and adjusted *p*-values (Benjamini-Hochberg correction) from the *t*-test performed are shown, as well as mean prenatal gene expression and mean postnatal gene expression.

Cluster id	Gene number	Prenatal expression	Postnatal expression	*p*-value	Adj. *p*-value
Turquoise	101	3.57653543	4.32510319	0.99999449	0.99999449

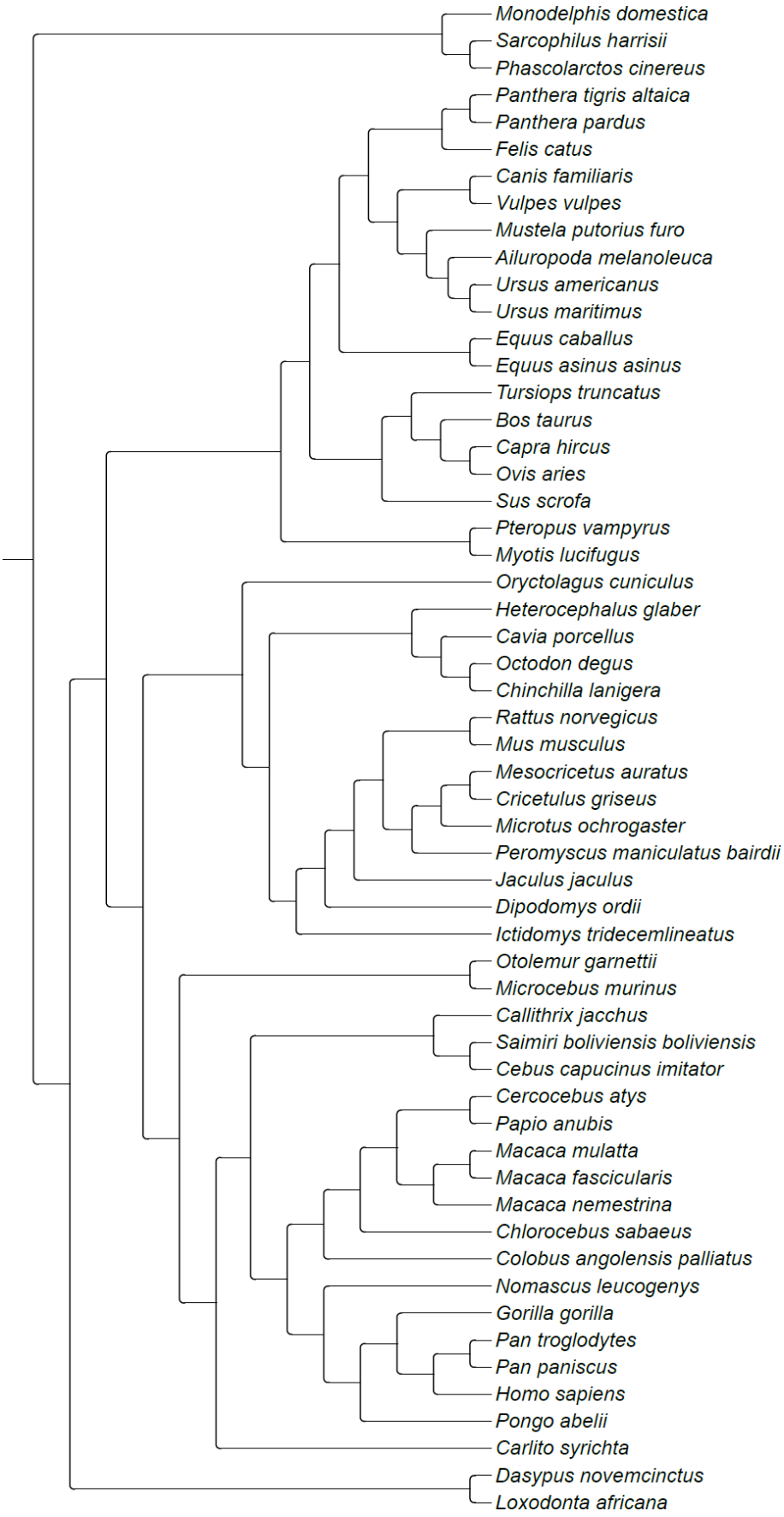


FIGURE S1. Phylogenetic tree of the set of species analysed. Distances obtained from the Timetree database (Kumar et al., 2017).

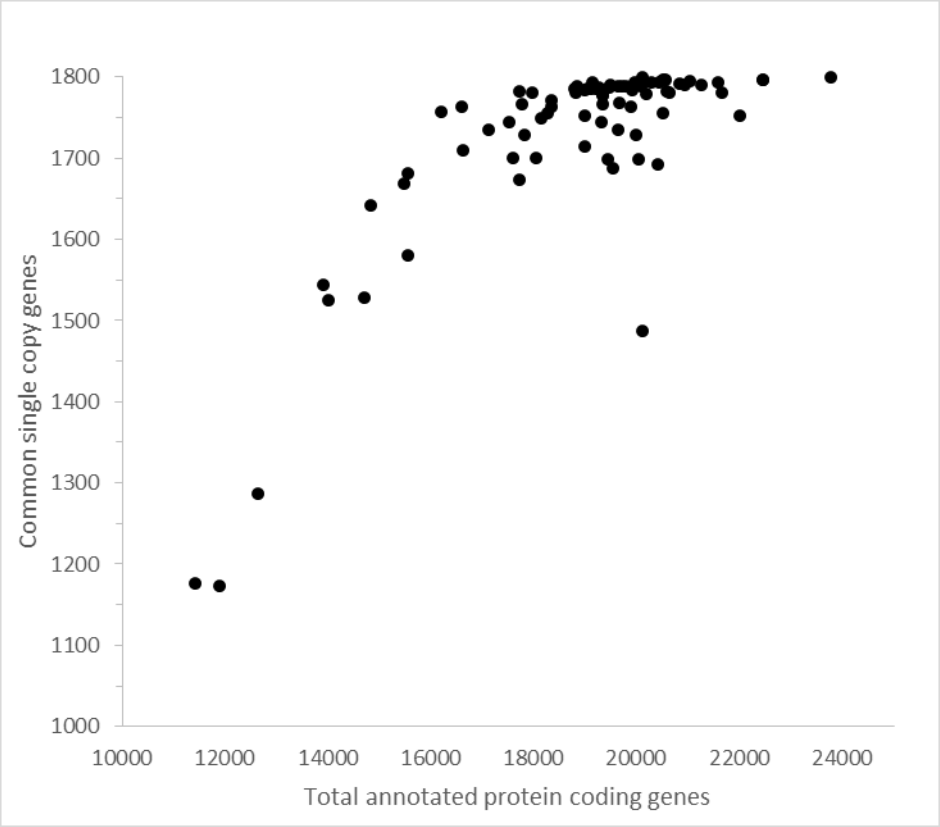
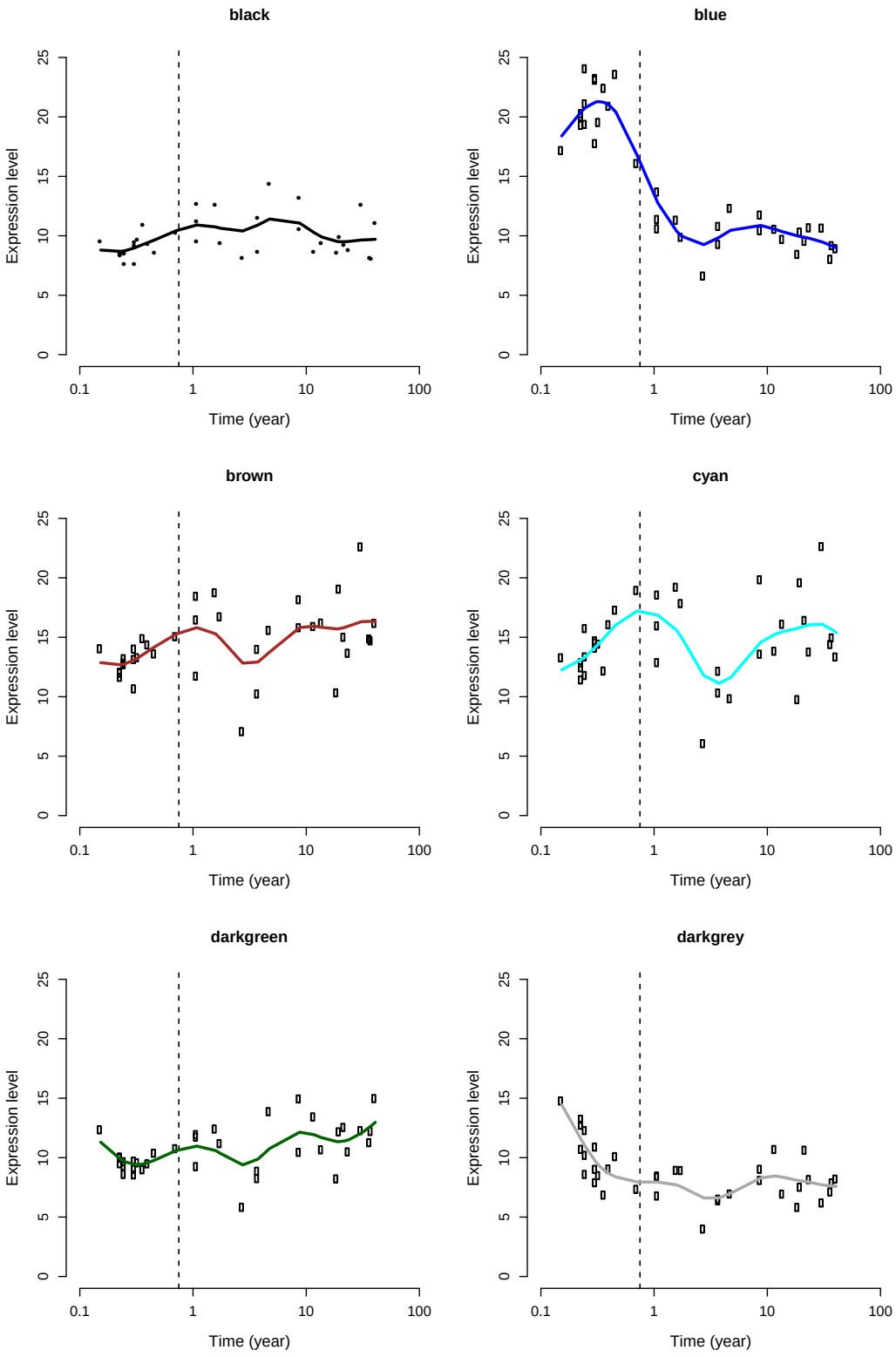
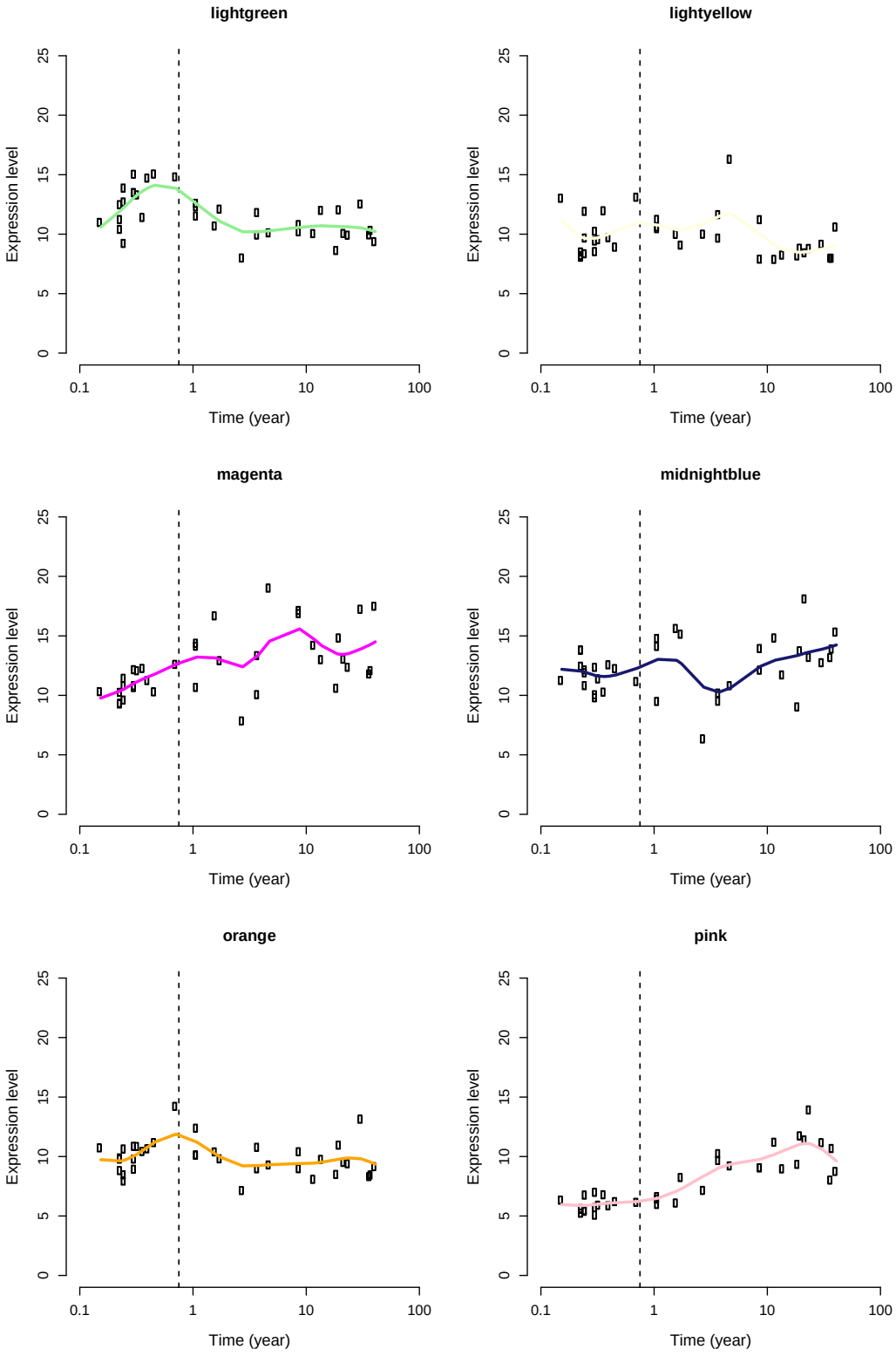
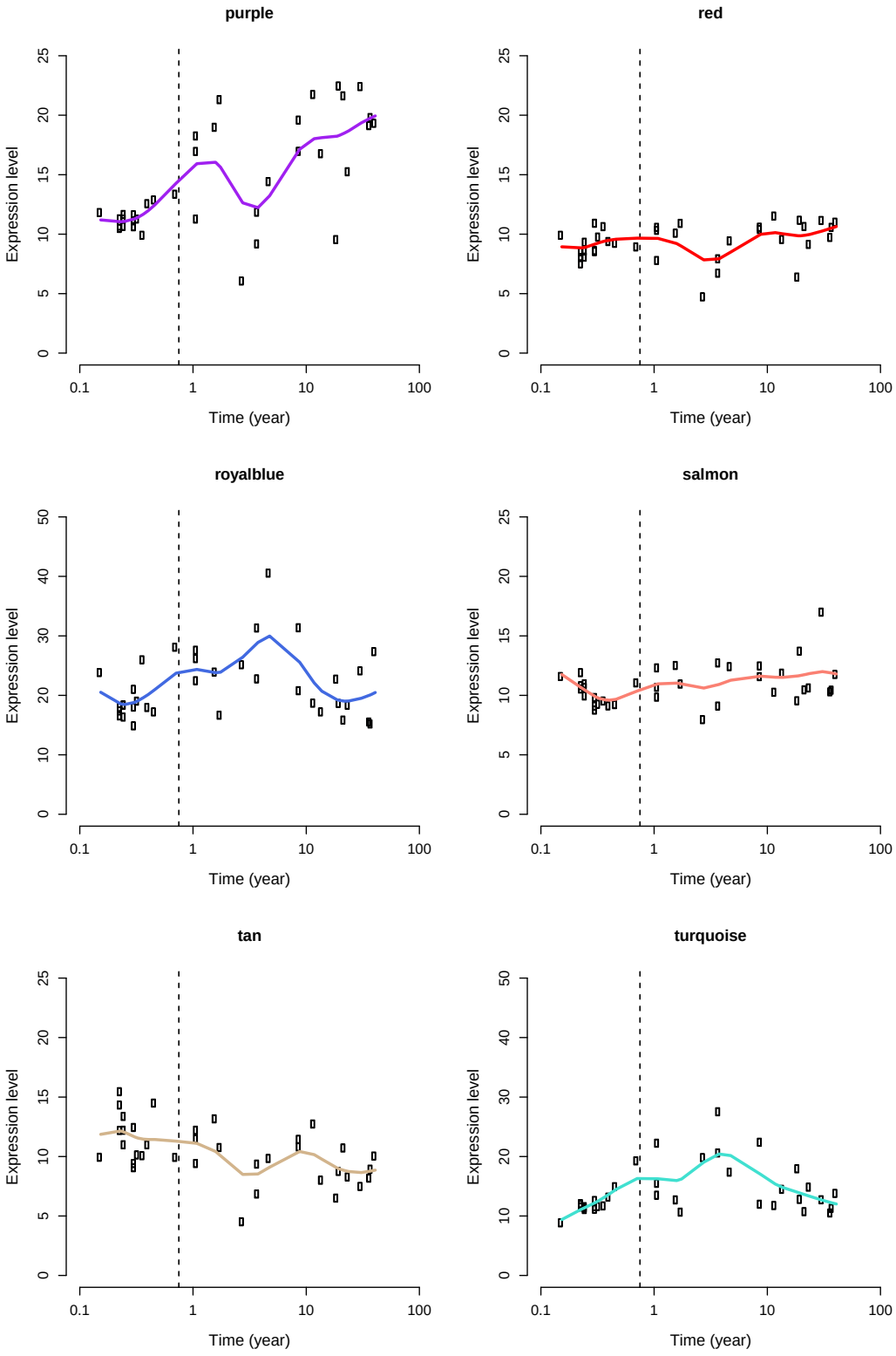


FIGURE S2. Plot of the total gene number of annotated genes per mammalian genome compared to the number of single copy core genes.


black
blue
brown
cyan
darkgreen
darkgrey
Expression level
Time (year)
0
5
10
15
20
25
0.1
1
10
100


lightgreen
lightyellow
magenta
midnightblue
orange
pink
Expression level
Time (year)
0
5
10
15
20
25
0.1
1
10
100


purple
red
royalblue
salmon
tan
turquoise
Expression level
Time (year)
0.1
1
10
100

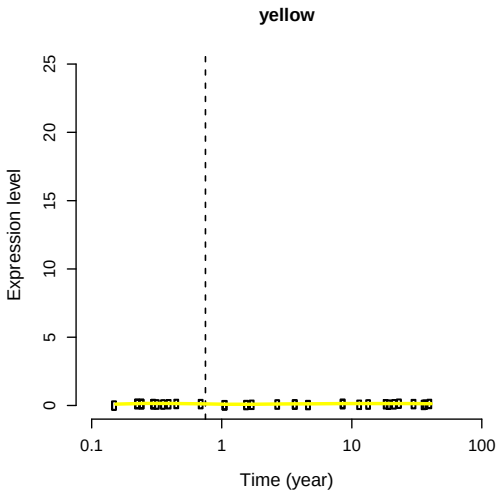
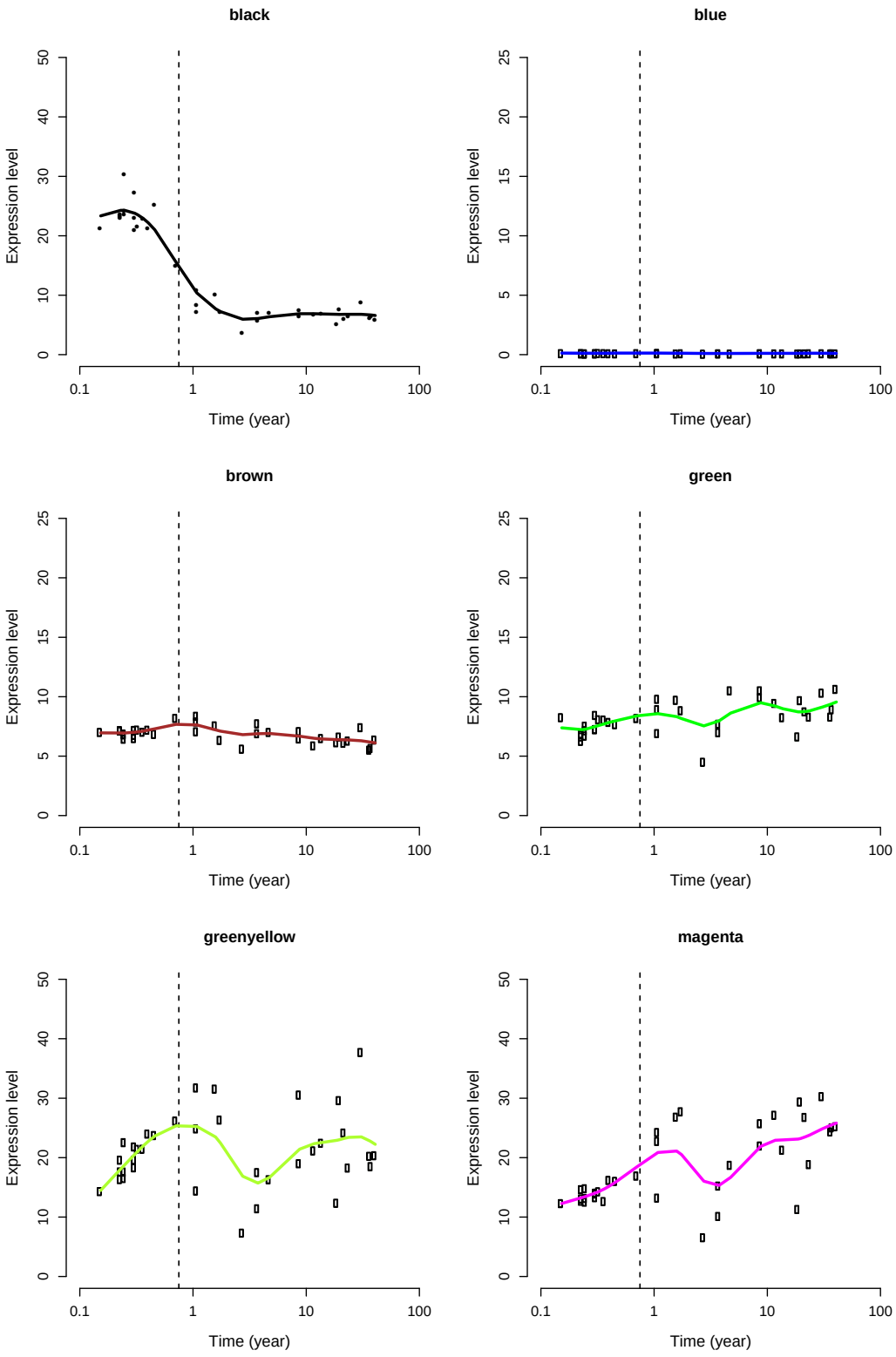


FIGURE S3. Gene expression across human development for each module obtained for encephalization associated genes. Birth point is indicated with a dashed line.


black
blue
brown
green
greenyellow
magenta
Expression level
Time (year)
0.1
1
10
100

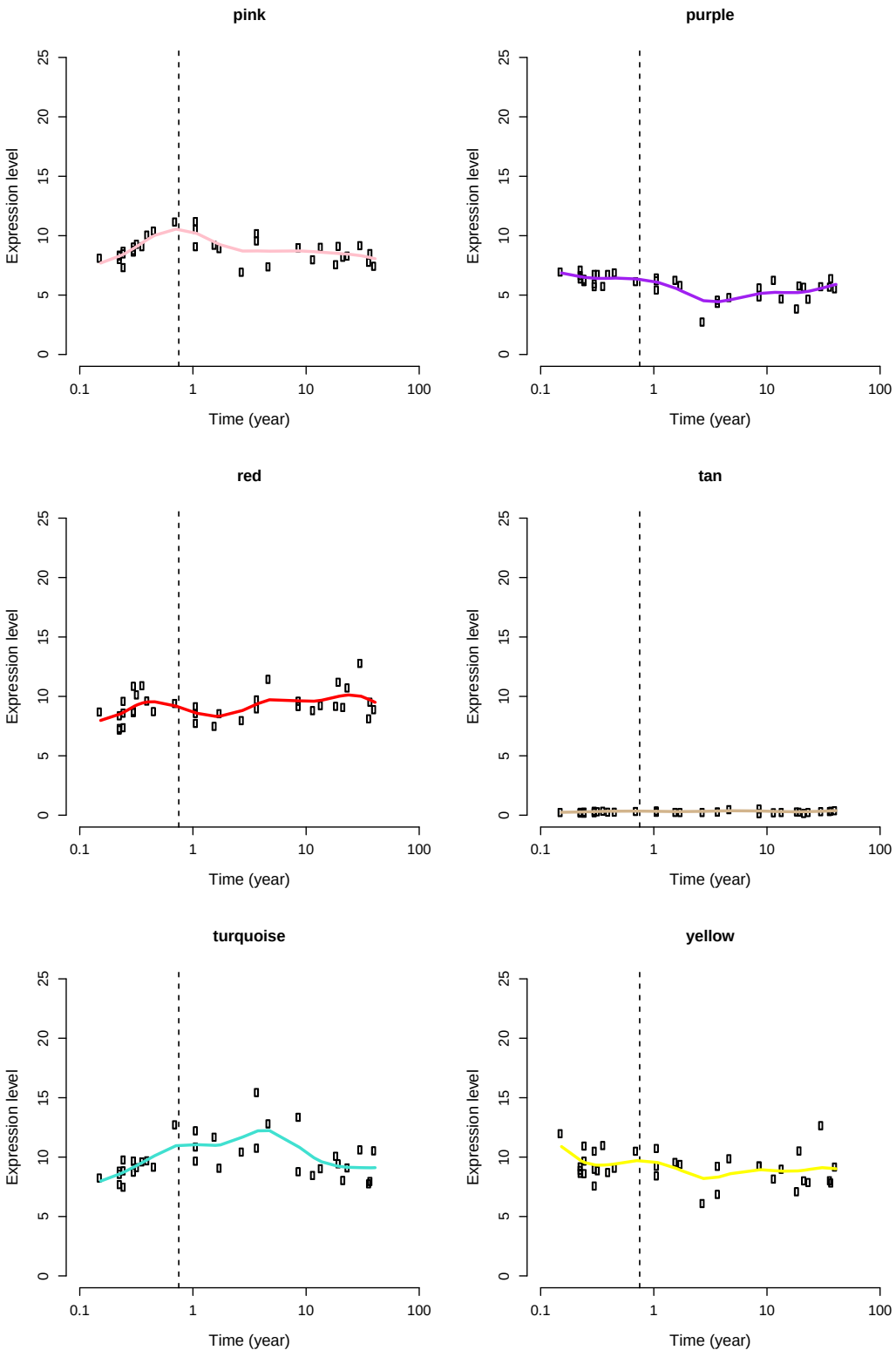


FIGURE S4. Gene expression across human development for each module obtained for neuron number associated genes. Birth point is indicated with a dashed line.

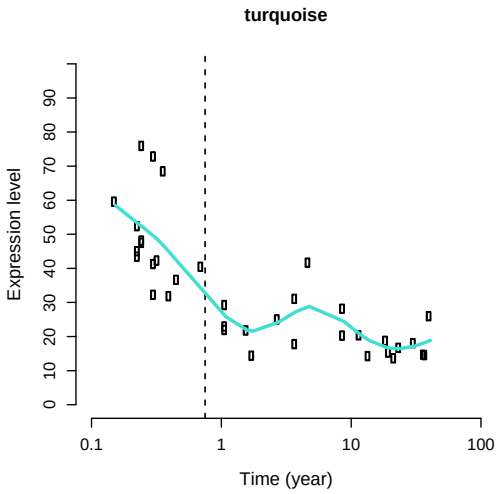


FIGURE S5. Gene expression across human development for each module obtained for Glia / neuron ratio associated genes. Birth point is indicated with a dashed line.

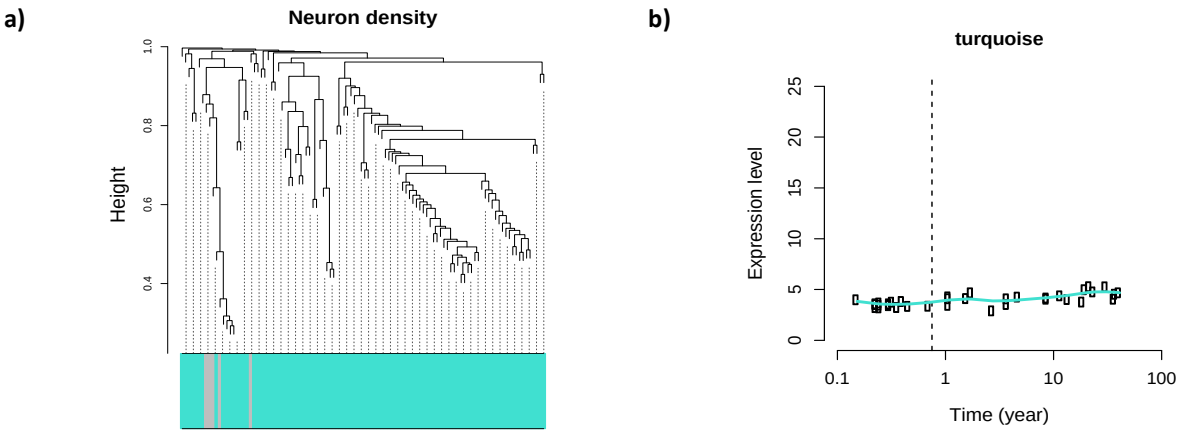


FIGURE S6. a) Gene cluster dendrogram and module colours for neuron density associated genes. **b)** Gene expression across human development for each module obtained. Birth point is indicated with a dashed line.

Chapter 3. Sexual size dimorphism is associated with variations in gene family size in gene families related to organism and neuronal development

Abstract

Sexual selection has a profound impact on species evolution, shaping their morphology, physiology and behaviour. In mammals, males often face higher competition to find mating partners compared to females and are often larger. Sexual size dimorphism –SSD; differences in body mass between males and females- is often taken as an indicator of the strength of sexual selection. Although the evolutionary implications of sexual selection have been a matter of intense research for many decades, the genomic signatures of sexual selection remain poorly understood. Gene family size evolution can reflect changes in the functional relevance of molecular pathways associated with the evolution of phenotypic traits. Here we used a comparative genomics approach to analyse gene family size (GFS) variations across 44 mammalian species for which annotated reference genomes and body mass estimates for both sexes are available. GFS variations were associated with SSD beyond chance expectations. Interestingly, gene families which have expanded in species with decreased dimorphism are enriched in a variety of functions related to brain development. No such associations were found for families with the greater expansion among species with the greatest dimorphism. These results are robust to phylogenetic correction and are not explained by covariance between sex size dimorphism and body mass. This is, to our best knowledge, the first systematic analysis of genome wide scale changes in gene family size and an indicator of sexual selection across the mammalian tree.

Key words: comparative genomics, gene family size, sexual selection, body mass, Rensch's rule

Main points

- We report a novel link between sexual size dimorphism, a commonly used indicator of the strength of sexual selection, in gene family size in mammalian species.

- The associations of gene family size and sexual size dimorphism are not explained by a known covariate factor, body mass, or phylogenetic relatedness.

- Expanded gene families in the least dimorphic species are enriched in functional categories related brain development.

- This is the first analysis exploring the links between gene family evolution and sexual selection.

Introduction

Sexual selection, which results from unequal competition between males and females for opportunities to contribute offspring to the next generation, plays a major role in evolution by influencing evolution and speciation (Darwin, 1901, Wilkinson et al., 2015). Sexual selection is considered to be of great evolutionary importance in determining a number of morphological and behavioural traits as well as long-term species viability. It can be distinguished from natural selection as it results from differences in relative mating success (Hosken and House, 2011). Sexual selection is thought to underlie a considerable amount of phenotypic diversity observed across taxa (Andersson, 1994). This type of selection can sometimes potentiate the effects of natural selection when the choosy sex prefers traits favoured by natural selection. However, sexual selection can also act in opposition natural selection on the same characteristic. For example, conspicuous ornamentation or coloration have been observed to provide an advantage in finding mating opportunities but at the same time increase the risk of attack by predators (Hernandez-Jimenez and Rios-Cardenas, 2012, Morgans et al., 2014, Husak et al., 2006, Håstad et al., 2005).

Sexual selection can drive the evolution of dimorphism between males and females in morphological, behavioural and physiological traits (McPherson and Chenoweth, 2012). Differences in size between the sexes can be observed in many taxa, sexual size dimorphism (SSD) refers to the differentiation in body size of sexually mature males and females within a species (Fairbairn, 1997). SSD results from selective pressures on body size acting disproportionally on one sex (Fairbairn, 1997, Dos Remedios et al., 2015, Andersson, 1994). For example, selection on fecundity can drive female body size increases as larger females can produce larger clutch sizes and larger offspring, resulting in larger females compared to males (Fairbairn, 1997, Baotic and Stoeger, 2017, Kingsolver and Pfennig, 2004, Serrano-Meneses et al., 2008). Among mammals and birds, however, male biased SSD is prevalent. A 10% or higher male size bias is observed in over

45% of mammal species (Fairbairn et al., 2007). Northern Elephant Seals are one of the starkest examples of sexual dimorphism in the animal kingdom, with a 1575kg average weight difference between males and females (Bininda-Emonds and L. Gittleman, 2000). Larger male sizes are attributable to greater intraspecific competition for mates and the success of larger males (Lindenfors et al., 2007).

SSD, in mammals, has been associated with higher overall species size. This association between sexual size dimorphism (SSD) with species body mass has been shown in many distinct groups of mammals (Weckerly, 1998, Kappeler et al., 2019, Lindenfors et al., 2007) whereby larger species tend to also be the ones with the highest male biased SSD. This is known as the Rensch's rule (Rudoy and Ribera, 2017, Fairbairn, 1997).

Although several hypotheses have been proposed to explain variations in body mass among species, sexual selection has been proposed to be a fundamental driver of body size increase (Blanckenhorn, 2000). A study examining the impact of various selective pressures on body mass revealed that higher sexual selection favouring larger male sizes is a major contributor to both SSD and body mass in a species (Dale et al., 2007). It is thought that competition between males favours increases in their body size. These increases in body size of males leads to, although at a slower pace, increases in female body size because of the physiological requirements of copulation with larger males and to permit the birth of larger male offspring (McLain, 1993, Lindenfors et al., 2007). In mammalian lineages sexual selection has been shown to be an important determinant of both sexual dimorphism as such, and of the general size increase (Lindenfors and S.Tullberg, 2011).

Several studies have focused on discussing the mechanisms that originate sex differences in size, (see review by (Fairbairn et al., 2007). Nevertheless, little is known about the molecular signatures associated with the evolution of SSD and accompanying morphological and behavioural traits. Changes in the developmental programmes for SSD and other co-evolving traits in species under higher or lower sexual selection are encoded in the genome of each species. Gene family size is a highly dynamic trait with a high proportion of gene families experiencing significant degrees of gene gain and loss (Demuth et al., 2006). Changes in gene family size (GFS) can provide insights into changes in the relative functional relevance of molecular functions and have been associated with a number of phenotypes (Castillo-Morales et al., 2014, Castillo-Morales et al., 2016, Niimura and Nei, 2005). Gene duplication has long been recognised as a major source of functional innovation in genomes (Holland et al., 2017, Tickle and Urrutia, 2017). It is an ideal focus for genomic basis of phenotypic evolution. Studies of gene family size variations are particularly useful when examining variation over large evolutionary timescales as it allows to examine gene families beyond highly conserved single copy genes in housekeeping pathways controlling the most basic cellular functions that would be amenable to sequence evolution analysis. Importantly, gene duplication of developmental related genes is considered to have played a major role in the evolution of several vertebrate features. The most clear-cut example is the duplication of the Hox gene clusters (Soshnikova et al., 2013, Wagner et al., 2003). Gene family size evolution has already been linked with the evolution of brain morphology and size (Castillo-Morales et al., 2014, Castillo-Morales et al., 2016).

Here, using a comparative genomics approach, we tested if changes in gene family size can be linked to body size dimorphism in 44 mammalian species for which SSD data is known. To ensure that any resulting associations between GFS and SSD are not a by-product of accompanying increases in body mass, we use comparative models including both variables.

Hypotheses

1. If variations in gene family size have contributed to the evolutionary changes associated with sexual selection, then we would expect to observe a higher number of associations between gene family size and body size dimorphism compared to null expectations.
2. It is well known that body mass dimorphism is related with overall body mass. If gene family size variations in line with SSD are not related to the relation with body mass, then we would expect to see that gene family associations will be robust to controlling for body mass.
3. If associated gene families underlie phenotypic evolution associated with variations in sexual selection, then we expect associated families to be enriched in developmental related functions.

Material and Methods

Sexual size dimorphism and body mass.

Body mass data for 44 mammalian species was collected from available literature, databases and institutional datasets (Table 1 and supplementary figure 1). Adult male and female body mass was used to calculate average body mass for the species. Male and female body mass was used to estimate sexual size dimorphism (SSD) values for each species. SSD measures for each species were calculated as the log2-transformed ratio of average male versus average female body mass (Dunham et al., 2013). For analyses using male and or female body mass log10 transformed values to make these variables normally distributed as most species tend to have relatively small body masses with a few species having large body masses.

Test for Rensch's rule

The Rensch's rule expects a positive relationship between male and female body mass but with female mass increasing at a lower pace in species with the largest males. Male and female body mass were log-transformed prior to analyses. Rensch's rule was tested using major axis regression (MA regression) or a geometric mean functional relationship as implemented in (Álvarez et al., 2013). MA regression expresses the slope describing the association between two variables in a symmetrical manner, regardless of which variable is taken as dependent and which as independent. MA regressions, crucially allow testing hypotheses relating to the slope (Fairbairn, 1997). Since Rensch's rule is detected when the allometric slope between males and females exceeds 1, we tested whether the resulting slope was different from 1 (unity) by using a procedure implemented in the R package "smatr" (slope.test) (see (Warton et al., 2006) for a description). A similar procedure was then used to compare the strength of associations between gene family size and with male and female body mass.

Gene family annotations

Gene family annotations were downloaded for 78 species of mammals from Ensembl-version 95 (Cunningham et al., 2018). A number of species were found to have an unusually low number of protein coding annotated genes which are likely to result from low sequencing coverage. To test if this observation resulted from low quality genomes rather than actual variations in gene number between species we identified a set of "single copy core" genes defined as those genes present in at least 90% of species and always as a single copy. Missing a high number of these single copy core genes is more likely to be explained by low quality genomes rather than actual variations in gene number. Consistent with this, we found that those genomes with under 16000 annotated protein coding genes were also missing a high number of the core gene set (Supplementary figure 2). The platypus was found to be an outlier in this distribution with a markedly low number of single copy core genes but with a *typical* number of total

annotated protein coding genes. This is likely to be explained by this distant mammalian species having a different gene set or highly diverged gene sequences rather than low sequencing coverage. Based on this, all species (10/78) with fewer than 16000 protein coding genes were not included in any analyses. This removed most of the association between the core gene set count and total number of annotated genes. No further correction for total gene number was carried out in line with most studies of gene family evolution. This is because, most gene families have one or two genes in each species and applying a normalisation factor would create an artificial between species variation in gene number in most families. We tested the association between total number of annotated genes and phenotype variation which showed no significant association for any phenotype tested.

Gene families in the 44 species with available body size dimorphism and a sequenced genome were required to have at least one gene in at least 80% of species ($n > 35$) to be included in analyses to remove newly evolved families only present in a specific branch. Only gene families with at least two genes in at least one species and a minimum difference of two between the maximum and the minimum number of genes across species were considered. This, removed gene families with no variation in gene number across species, single copy genes only varying in their presence/absence and families where all variation is explained by a single deletion or duplication event. After applying these filters, 6036 families were included in the study.

Phylogenetic regressions of gene family size and cell composition parameters

We used generalised least square regressions to assess the strength of associations between gene family size and phenotype variations across species. To rule out associations between phenotypic parameters and variations in GFS being explained by shared ancestry, a phylogenetically generalised least squares regression (PGLS) (Grafen, 1989, Grafen, 1992)

was used. PGLS is a widely used method for assessing the association between variables in a set of species and involves constructing the phylogenetic variance-covariance matrix taking into account the phylogenetic tree which is then used to perform a generalized least squares linear regression. Phylogenetically corrected regressions were performed using the "nlme" R package (Pinheiro et al., 2018) assuming a Brownian motion model of evolution and a maximum likelihood method.

Some families were found to be associated to more than one phenotype. To account for this, functional enrichments in sets of gene families were reassessed after including all phenotypes tested in a single PGLS model.

Effect size

Cohen's *r* effect sizes (Cohen, 1988) were calculated by computing correlation r from the *t* statistic of the PGLS model summary using the following formula (Rosenthal et al., 2000):

$$r = \frac{t}{\sqrt{t^2 + df}}$$

Where total degrees of freedom are calculated as the number of degrees of freedom in the model -total number of variables (one to two) plus the intercept minus one- subtracted from the total degrees of freedom -sample size (number of species in the test) minus one. Focusing on effect sizes, instead of on *p* values reduces the probability of type two errors, where an alternative hypothesis would be wrongly rejected, particularly when sample sizes are low (Nakagawa, 2004) as is the case in this study (n = 44) .

Power analysis

Power analysis was carried out to calculate the minimum number of species required to have to achieve the recommended statistical power of 0.8 (Cohen, 1988) to assess significance of medium effect sizes (r > 0.3)

(Cohen, 1988) when testing associations for 6036 gene families.

For this we used the following formula from (Cummings and Hulley, 1988a) as implemented in (http://www.sample-size.net/correlation-sample-size/):

$$N = \left[\frac{Z_\alpha + Z_\beta}{C}\right]^2 + 3$$

Where Z_α is the standard normal deviate for the significance threshold 0.05 divided by the number of tests carried out; Z_β is the normal deviate of the accepted level of type two errors (0.2 for a statistical power of 0.8) and C is calculated as follows:

$$C = 0.5 * \ln\left[\frac{1+r}{1-r}\right]$$

Where r refers to the size of association for which significance should be reliably established.

Effective number of tests

The effective number of tests was calculated to take into account the non-independence and collinearity between individual data points in a sample to calculate the 'true' sample size to avoid inflating the number of tests when carrying out multiple testing corrections. This was calculated from the eigenvalues from a correlation matrix of gene family size profiles (Li and Ji, 2005).

Effect size distribution randomisation test

To test if the number of gene families associated with each phenotype after applying these thresholds is higher than random expectations, the number of gene families meeting this threshold were compared against the number reaching this threshold in 1,000 randomised data sets. In each randomisation, measurements for each phenotype were randomly re-assigned to species name, keeping gene family size data for each species unchanged. This test was carried out using absolute values of r or signed

values. In a one tail test, if fewer than 50 of the randomised distributions had a larger number of families with a large effect size than the real distribution of *r*s, then the test was deemed to be significant (alpha value of 0.05 in a one-tail test).

Gene ontology term enrichment analysis

Gene ontology (GO) functional terms annotations for each gene for each species were obtained from the Gene Ontology Consortium database (www.geneontology.org). GO terms were linked to a family whenever that term was assigned to any gene in the family in any of the 78 sequenced mammalian species available in Ensembl version 95 (Cunningham et al., 2018). GO terms associated with fewer than 50 associated gene families were pooled together into a single category labelled “small GO” as overrepresentation of categories associated with very few genes would be difficult to assess and would unnecessarily reduce statistical power (Castillo-Morales et al., 2014). Unlike the case in other studies, families not associated with any functional GO term were included in the analyses under an “unknown GO” term. Enrichment of GO categories among the set of gene families associated to each of the phenotypes of interest, was carried out by measuring the proportion of families assigned to each GO term within the analysed set of gene families and comparing it with the proportion of gene families associated to each GO term in 1,000 equally-sized samples of randomly chosen gene families from the background set. The mean and standard deviation of GO term representation as measured in each of these 1000 random samples were taken to determine the corresponding *p*-values for each GO term using *Z*-score with the formula described in the above methods section and Benjamini-Hochberg correction for multiple testing as implemented in Castillo-Morales et al. (2014).

Results

In order to assess the relationship between SSD and changes in GFS, correlation coefficients between SSD and GFS were calculated for each gene family in 44 mammalian species (Table 1, Supplementary Figure 1). Overall, associations with SSD were not as strong as those observed in the previous chapter examining brain cell composition. Applying a cut-of value of r = 0.3, for medium size effect (Cohen, 1988) as very few families had large effect size associations, we observed a total of 243 positively associated gene families and 580 of negatively associated gene families (Figure 1a and b). According to power calculations, a total of 96 species would be required to have enough statistical power to assess significance of large effect size associations and 296 species for medium effect size associations. As expected, no gene families were found to be significantly associated with SSD after correcting for multiple testing (adj. p > 0.05). Calculating the effective number of tests (n = 5823), does not reduce the number of species that would be required to have statistical power for testing significance of individual gene families (n = 96 for large effect sizes and n = 295 for medium effect sizes).

However, when examining the overall distribution of r values, we found that the distribution of correlation coefficients showed a higher dispersion than that expected by chance, with a higher number of associations with an absolute value for r > 0.3 compared to correlations of randomised data (p = 0.012). No significant deviations of the overall distribution or excess of medium and or large effect size associations were observed in one tail of the distribution were observed when examining the distributions, with no excess of positive or negative associations compared to random distributions observed (p > 0.05). This result suggests that there is an excess of gene families which have preferentially expanded in species at either extreme of the SSD distribution compared to chance expectations.

Examining the sets of gene families most strongly associated with high SSD

(r > 0.3), which are families which are larger in species with the highest male biased size dimorphism, gene ontology enrichment analysis revealed significant enrichment in regulation of cell adhesion and stimulatory C-type lectin receptor signalling pathway in the high SSD associated families. Examining those families which have expanded in the least dimorphic species (r < -0.3) revealed significant overrepresentation of functional categories related to various aspects of brain development (Figure 2a).

SSD has been previously associated with body mass in several taxa including mammalian groups. Thus, we tested the association between SSD and body mass in the set of species examined. According to the Rensch's rule, species with the highest levels of SSD also are the largest overall (Abouheif and Fairbairn, 1997). We found that the Rensch's rule does apply in the species under study; as larger male sizes are associated with increasing departure from one to one relation in size with females (slope = 1.03, upper – lower confidence intervals: 1.017 – 1.053, p > 0.001; Figure 3). Consistent with this, we found a significant and positive association between SSD and body mass (r = 0.392; p = 0.009). If male body mass is under higher selective pressure with female body mass mostly evolving to catch up with male mass, then we would expect that, overall, gene family size variations should correlated more strongly with male than to female mass. Indeed, we find that this is the case (slope = 1.03, upper – lower confidence intervals: 1.027 – 1.034, p > 0.001; Figure 4).

To examine whether the observed associations and functional enrichments were not the result of covariance between SSD and body mass, body mass was included in a model containing SSD. A total of 326 gene families were found to have positive associations of medium effect sizes with SSD and 403 with negative associations. Enrichment analysis of functional categories on SSD associated families was recalculated after controlling for the contribution of body mass. We identified significant enrichments among SSD positively associated families (those expanding in the most dimorphic

species) in the stimulatory C-type lectin receptor signalling pathway. Among gene families expanding in the least dimorphic species, we found an overrepresentation of functional categories linked to brain development (Figure 2b).

Discussion

The present study aimed to explore the genomic signatures associated with sexual selection in mammalian species. Higher sexual selection is associated with higher levels of dimorphism between males and females. Body size is a commonly dimorphic trait and has been taken as an indicator of the strength of sexual selection. Here we assessed the link between changes in gene family size and variations in sexual size dimorphism (measured as dimorphism for overall body size) in 44 mammalian species correcting for phylogenetic relatedness.

We found stronger than expected associations between gene family size and body size dimorphism which is consistent with changes in gene family size contributing to the evolution of this trait as expected under hypothesis one. Generally, stronger associations between gene family size and male body mass than to female body mass which is consistent male body mass being the primary target of selective pressure in the set of species examined.

Functional annotation enrichment analysis among gene families most strongly associated with SSD showed enrichment of biological functions related to regulation of cell adhesion and stimulatory C-type lectin receptor (CLRs) signalling pathway. The CLR family of proteins is involved in cell adhesion and immune response through signalling cascades that induce the production inflammatory mediators that coordinate the elimination of

pathogens and infected cells (Takeuchi and Akira, 2010). Interestingly, families most expanded among the least dimorphic species show significant enrichments in a variety of functional categories associated with brain development. The fact that a set of functional categories are enriched among gene families associated with SSD is consistent with expectations under hypothesis three.

SSD has been linked to overall species body mass in several lineages (Weckerly, 1998, Kappeler et al., 2019, Lindenfors et al., 2007, Jiménez-Arcos et al., 2017). The Rensch's rule pattern is observed when increases in male mass correlated with smaller increases in female mass produce a tendency for size dimorphism to scale with body size (Abouheif and Fairbairn, 1997). In our set of species we found support for the Rensch's rule with female body mass increases lagging further from males in those species with the largest males. This results provide an indication that there exists parallel, though not equal, selection pressures on males and females (Lindenfors et al., 2007). After controlling for the contribution of body mass in a PGLS model containing both variables, we found that associations with brain developmental functions remain significant for SSD. This is consistent with expectations under hypothesis two. This is suggestive of a complex interplay in the evolution of sexual selection, body mass and brain complexity. Sexual size dimorphism is suggested to have evolved in polygynous species with larger body mass being favoured when males fight over mating access to females (Pérez-Barbería et al., 2002). In contrast, lower rates of body size dimorphism is primarily seen in species with monogamous mating systems and often, with bi-parental care for offspring (Weckerly, 1998, Pérez-Barbería et al., 2002, Kleiman, 1977). It is likely that such close social bonds require increasingly complex social skills leading to changes in brain circuitry.

Despite the recognized importance of sexual selection and brain evolution variation among species, few studies have directly approached the relationship between these two features (Schillaci, 2006, Pitnick et al., 2006, Garamszegi et al., 2005, Madden, 2001). Research in avian species has led to propose that sexual selection drives brain size dimorphism (Madden, 2001, Garamszegi et al., 2005). In primates it has been shown that the largest relative brain sizes are associated with monogamous mating systems, leading to suggest that primate monogamy requires greater social acuity and the ability to manipulate others within the group (Schillaci, 2006). Moreover, among primate species it was also observed that increasing levels of body mass dimorphism are associated with decreasing relative brain size (Schillaci, 2006). Our results suggesting an expansion of brain development related families among less dimorphic species provide the first molecular support to these phenotypic associations.

Future analyses should also expand on the interplay of sexual selection in brain evolution. Given that the differing demands between males and females should be expected to produce variation in the relative sizes of various brain structures (Lindenfors and S.Tullberg, 2011), it would be of great interest to analyse data on brain structures for the sexes separately.

With testes size relative to body mass having also been proposed to be an indicator of sexual selection, in future studies it would be interesting to analyse the associations of gene family size and this additional indicator of sexual selection. It is further possible that the associations between SSD and brain development might also apply to relative testis size. A comparative analysis of brain, testis, and social and mating systems data for more than 300 Chiroptera species showed that species with promiscuous females are associated with smaller brains and larger testes, while species with females exhibiting mate fidelity are associated with significantly larger brains and smaller testicles (Pitnick et al., 2006). These results suggest an evolutionary

trade-off between investment in testes and the development of brain size due to the metabolic cost of both tissues (Pitnick et al., 2006).

In summary, we find significant shifts in gene family size associated with body size dimorphism and body mass which have both been associated with the strength of sexual selection. The most strongly associated gene families to sex size dimorphism and which have expanded in species with decreased dimorphism are enriched in a variety of functions related with brain development. No such associations were found for families with the greater expansion among species with the greatest dimorphism. These results are robust to phylogenetic correction and are not explained by covariance between sex size dimorphism and body mass. To our knowledge, this is the first study to suggest an association at the molecular level between decreased sexual selection.

Tables

TABLE 1. Data for sexual size dimorphism and body mass for mammalian species with fully sequenced reference genome available.

Common name	Species name	Male body mass, g	Female body mass, g	SSD	Average body mass, g
Algerian mouse	*Mus spretus*	14.552 (1)	13.69 (1)	0.09	14.1
American black bear	*Ursus americanus*	86000 (2)	54050 (2)	0.67	70025
Bonobo	*Pan paniscus*	39000 (3)	31000 (3)	0.33	35000
Cat	*Felis catus*	3600 (4)	2800 (4)	0.36	3200
Chimpanzee	*Pan troglodytes*	48900 (5)	40600 (5)	0.27	44750
Cow	*Bos taurus*	500000 (6)	425000 (6)	0.23	462500
Crab-eating macaque	*Macaca fascicularis*	5335 (2)	3588 (2)	0.57	4461.5
Degu	*Octodon degus*	258 (7)	232 (7)	0.15	245
Dog	*Canis familiaris*	29483.5 (8)	27215.5 (8)	0.12	28349.5
Dolphin	*Tursiops truncatus*	259000 (9)	194400 (9)	0.41	226700
Donkey	*Equus asinus asinus*	142000 (10)	141000 (10)	0.01	141500
Drill	*Mandrillus leucophaeus*	17000 (3)	10000 (3)	0.77	13500
Elephant	*Loxodonta africana*	5314000 (3)	2986000 (3)	0.83	4150000
Ferret	*Mustela putorius furo*	1991 (11)	954 (11)	1.06	1472.5
Goat	*Capra hircus*	67143 (6)	46429 (6)	0.53	56786
Golden Hamster	*Mesocricetus auratus*	121.6 (12)	107.4 (12)	0.18	114.5
Gorilla	*Gorilla gorilla*	160000 (5)	93000 (5)	0.78	126500.0
Guinea Pig	*Cavia porcellus*	813.3 (5)	630.4 (5)	0.37	721.9
Horse	*Equus caballus*	407000 (6)	358333 (6)	0.18	382666.5
Human	*Homo sapiens*	70000 (5)	60000 (5)	0.22	65000
Kangaroo rat	*Dipodomys ordii*	52.42 (5)	55.6 (5)	-0.08	54
Koala	*Phascolarctos cinereus*	10400 (13)	8200 (13)	0.34	9300
Long tailed chinchilla	*Chinchilla lanigera*	417.15 (2)	422 (2)	-0.02	419.6

Common name	Species name	Male body mass, g	Female body mass, g	SSD	Average body mass, g
Macaque	*Macaca mulatta*	9159.5 (2)	5339 (2)	0.78	7249.3
Marmoset	*Callithrix jacchus*	317.9 (14)	322 (14)	-0.02	320
Microbat	*Myotis lucifugus*	7.67 (15)	8.44 (15)	-0.14	8.1
Mouse	*Mus musculus*	20.3 (5)	17.1 (5)	0.25	18.7
Mouse Lemur	*Microcebus murinus*	86 (3)	84 (3)	0.03	85
Northern American deer mouse	*Peromyscus maniculatus bairdii*	15.9 (16)	17.2 (16)	-0.11	16.6
Olive baboon	*Papio anubis*	22957 (2)	13924.5 (2)	0.72	18440.8
Opossum	*Monodelphis domestica*	100 (2)	90 (2)	0.15	95
Panda	*Ailuropoda melanoleuca*	117000 (17)	97000 (17)	0.27	107000
Pig	*Sus scrofa*	90000 (6)	51700 (6)	0.80	70850
Pig-tailed macaque	*Macaca nemestrina*	10000 (3)	7000 (3)	0.51	8500
Polar bear	*Ursus maritimus*	465000 (18)	194100 (18)	1.26	329550
Prairie vole	*Microtus ochrogaster*	41.6 (19)	41.7 (19)	0.00	41.7
Rabbit	*Oryctolagus cuniculus*	1385 (5)	1381 (5)	0.00	1383
Rat	*Rattus norvegicus*	325.3 (5)	294.7 (5)	0.14	310
Red fox	*Vulpes vulpes*	6300 (20)	5300 (20)	0.25	5800
Sheep	*Ovis aries*	26000 (3)	20000 (3)	0.38	23000
Sooty mangabey	*Cercocebus atys*	11000 (3)	5600 (3)	0.97	8300
Tasmanian devil	*Sarcophilus harrisii*	6750 (5)	5000 (5)	0.43	5875
Tiger	*Panthera tigris altaica*	227500 (21)	147500 (21)	0.63	187500
Vervet-AGM	*Chlorocebus sabaeus*	5881.5 (2)	4218 (2)	0.48	5049.8

SSD = Sexual size dimorphism.

(1) (Nunes et al., 2001). (2) (Research, 2016a). (3) (Weckerly, 1998). (4) (Yamane et al., 1996). (5) (Jones et al., 2009). (6) (Mysterud, 2000). (7) (Suarez and Mpodozis, 2009). (8) (Société des Produits Nestlé S.A., 2019). (9) (Read et al., 1993). (10) (Nengomasha et al., 2016). (11) (He et al., 2002). (12) (Gattermann et al., 2002). (13) (Nowak and Dickman, 2005). (14) (Araújo et al., 2000). (15) (Kurta and Kunz, 1988). (16) (Wolff, 1985). (17) (Charlton et al., 2009). (18) (Kingsley, 1979). (19) (Bondrup-Nielsen and Ims, 1990). (20) (Macdonald and Sillero-Zubiri, 2004). (21) (Lerner, 2013).

Figure legends

FIGURE 1. Gene family associations with sexual size dimorphism and body mass in mammals. Venn diagram comparing gene families displaying a **a)** positive and **b)** negative association between GFS and each phenotype tested: sexual size dimorphism and body mass. Ovals are not drawn to scale.

FIGURE 2. Functional annotation overrepresentation among gene families associated with each phenotype. Heatmaps show gene ontology (GO) biological process category enrichments among members of gene families associated **a)** positively and **b)** negatively with sexual size dimorphism and body mass in mammals. Only significantly overrepresented terms in at least one model (with a cut off threshold for significance of 0.05 and using Benjamini-Hochberg FDR correction for multiple testing) are shown.

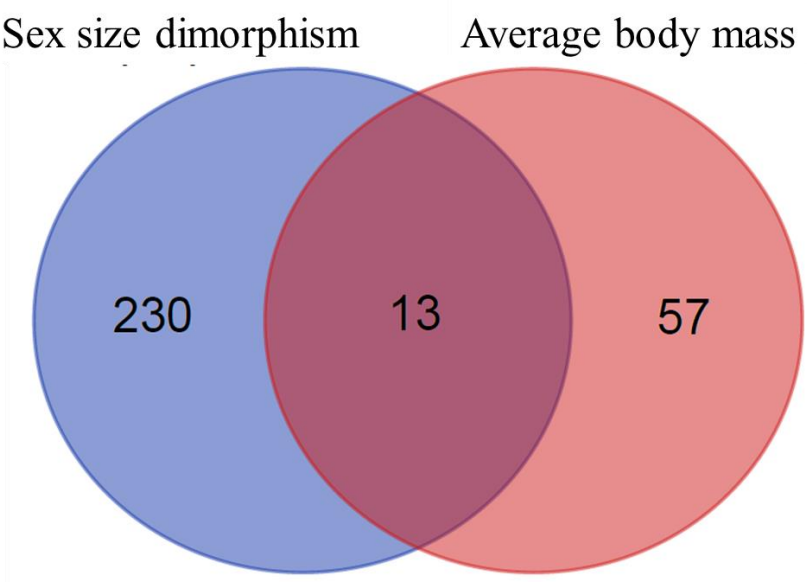
FIGURE 3. Rensch's rule in mammals. The calculated slope between male and female body size (dashed line; MA regression) is significantly different from unity (continuous line) (slope = 1.03, upper – lower confidence intervals: 1.017 – 1.053, $p > 0.001$).

FIGURE 4. Comparison of correlation coefficients for gene family size with male and female mass (dashed line; MA regression). The calculated slope for male and female *r* values was found to be significantly different from unity (continuous line) (slope = 1.03, upper – lower confidence intervals: 1.027 – 1.034, $p > 0.001$).

Figures

FIGURE 1.

a)



b)

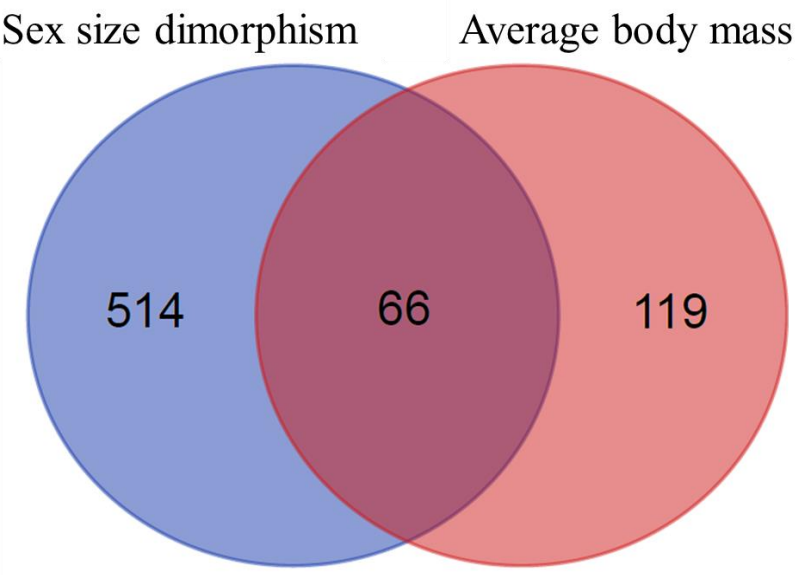
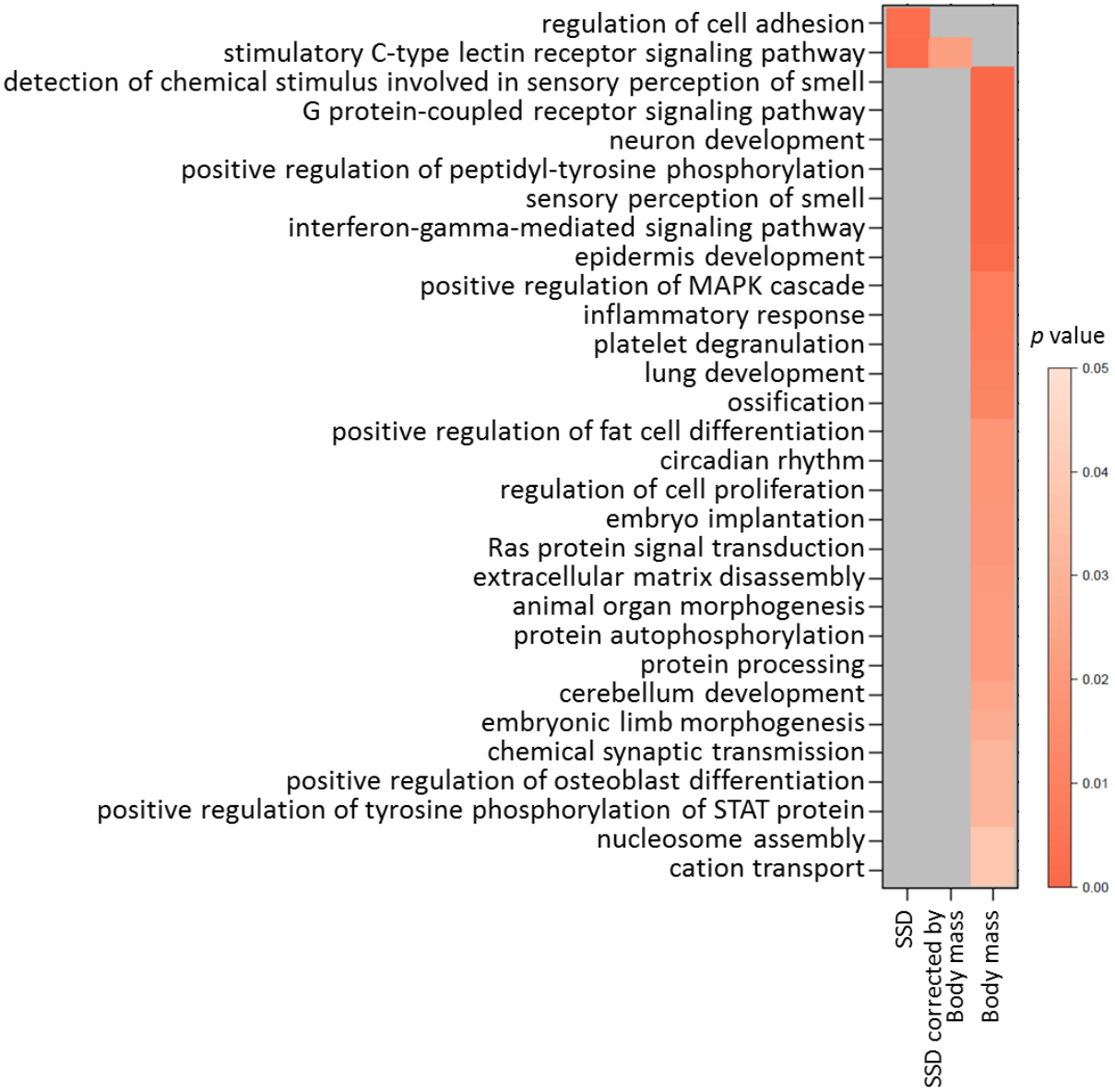


FIGURE 2.

a)


regulation of cell adhesion
stimulatory C-type lectin receptor signaling pathway
detection of chemical stimulus involved in sensory perception of smell
G protein-coupled receptor signaling pathway
neuron development
positive regulation of peptidyl-tyrosine phosphorylation
sensory perception of smell
interferon-gamma-mediated signaling pathway
epidermis development
positive regulation of MAPK cascade
inflammatory response
platelet degranulation
lung development
ossification
positive regulation of fat cell differentiation
circadian rhythm
regulation of cell proliferation
embryo implantation
Ras protein signal transduction
extracellular matrix disassembly
animal organ morphogenesis
protein autophosphorylation
protein processing
cerebellum development
embryonic limb morphogenesis
chemical synaptic transmission
positive regulation of osteoblast differentiation
positive regulation of tyrosine phosphorylation of STAT protein
nucleosome assembly
cation transport
SSD
SSD corrected by Body mass
Body mass
p value
0.05
0.04
0.03
0.02
0.01
0.00

b)

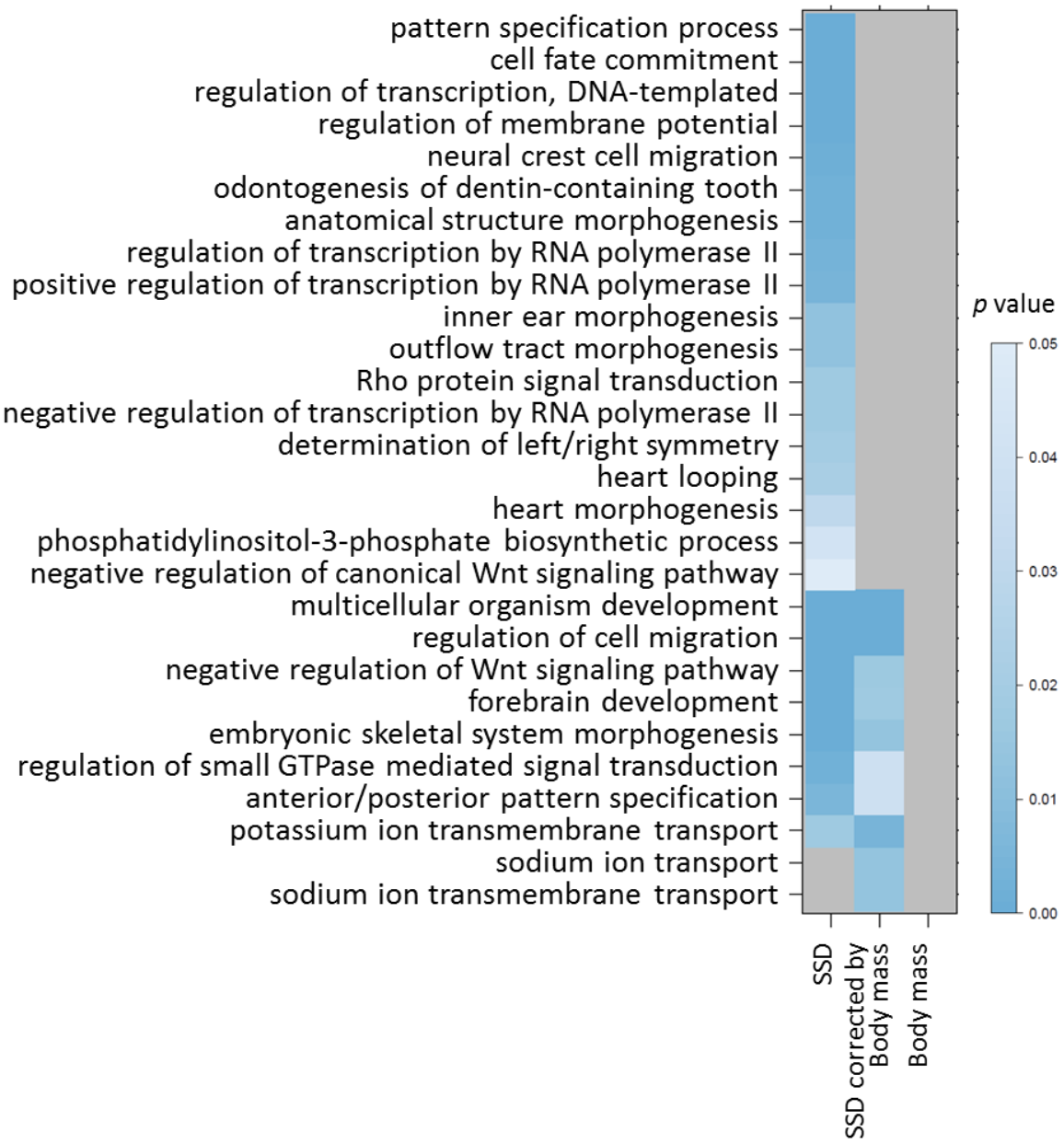

pattern specification process
cell fate commitment
regulation of transcription, DNA-templated
regulation of membrane potential
neural crest cell migration
odontogenesis of dentin-containing tooth
anatomical structure morphogenesis
regulation of transcription by RNA polymerase II
positive regulation of transcription by RNA polymerase II
inner ear morphogenesis
outflow tract morphogenesis
Rho protein signal transduction
negative regulation of transcription by RNA polymerase II
determination of left/right symmetry
heart looping
heart morphogenesis
phosphatidylinositol-3-phosphate biosynthetic process
negative regulation of canonical Wnt signaling pathway
multicellular organism development
regulation of cell migration
negative regulation of Wnt signaling pathway
forebrain development
embryonic skeletal system morphogenesis
regulation of small GTPase mediated signal transduction
anterior/posterior pattern specification
potassium ion transmembrane transport
sodium ion transport
sodium ion transmembrane transport
SSD
SSD corrected by Body mass
Body mass
p value
0.05
0.04
0.03
0.02
0.01
0.00

FIGURE 3.

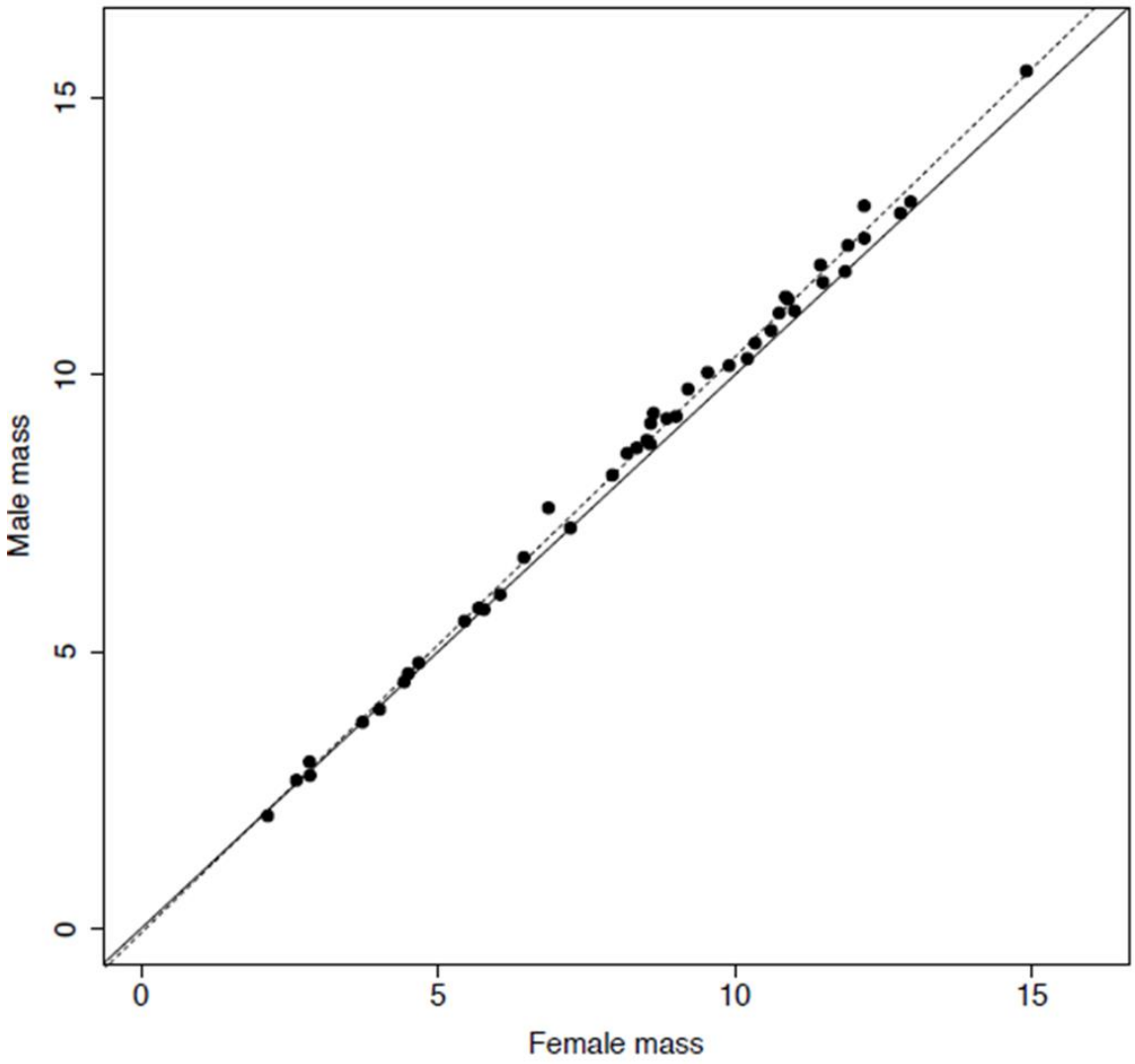
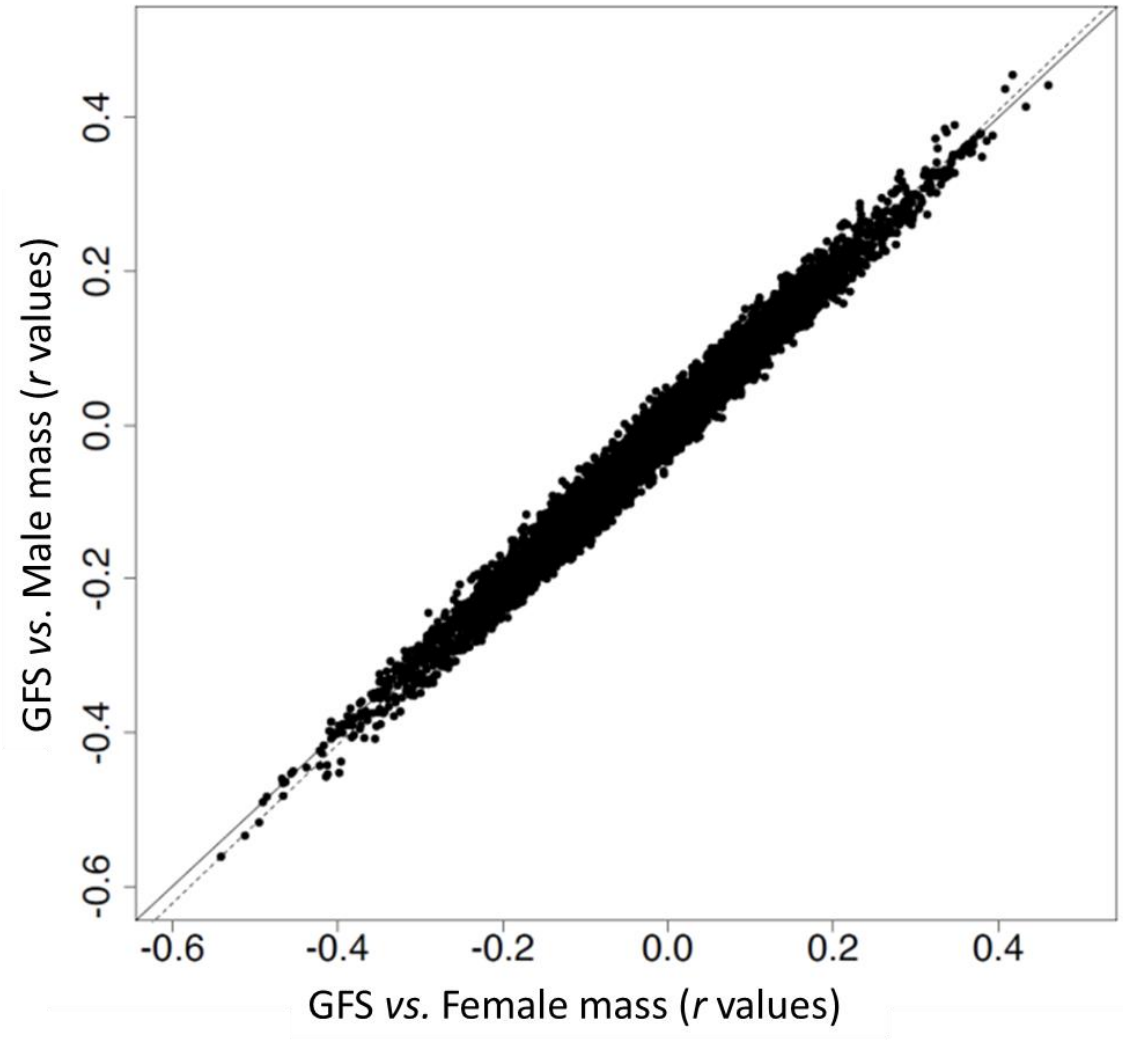


FIGURE 4.


GFS *vs.* Male mass (*r* values)
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.6
-0.4
-0.2
0.0
0.2
0.4
GFS *vs.* Female mass (*r* values)

Supplementary Material

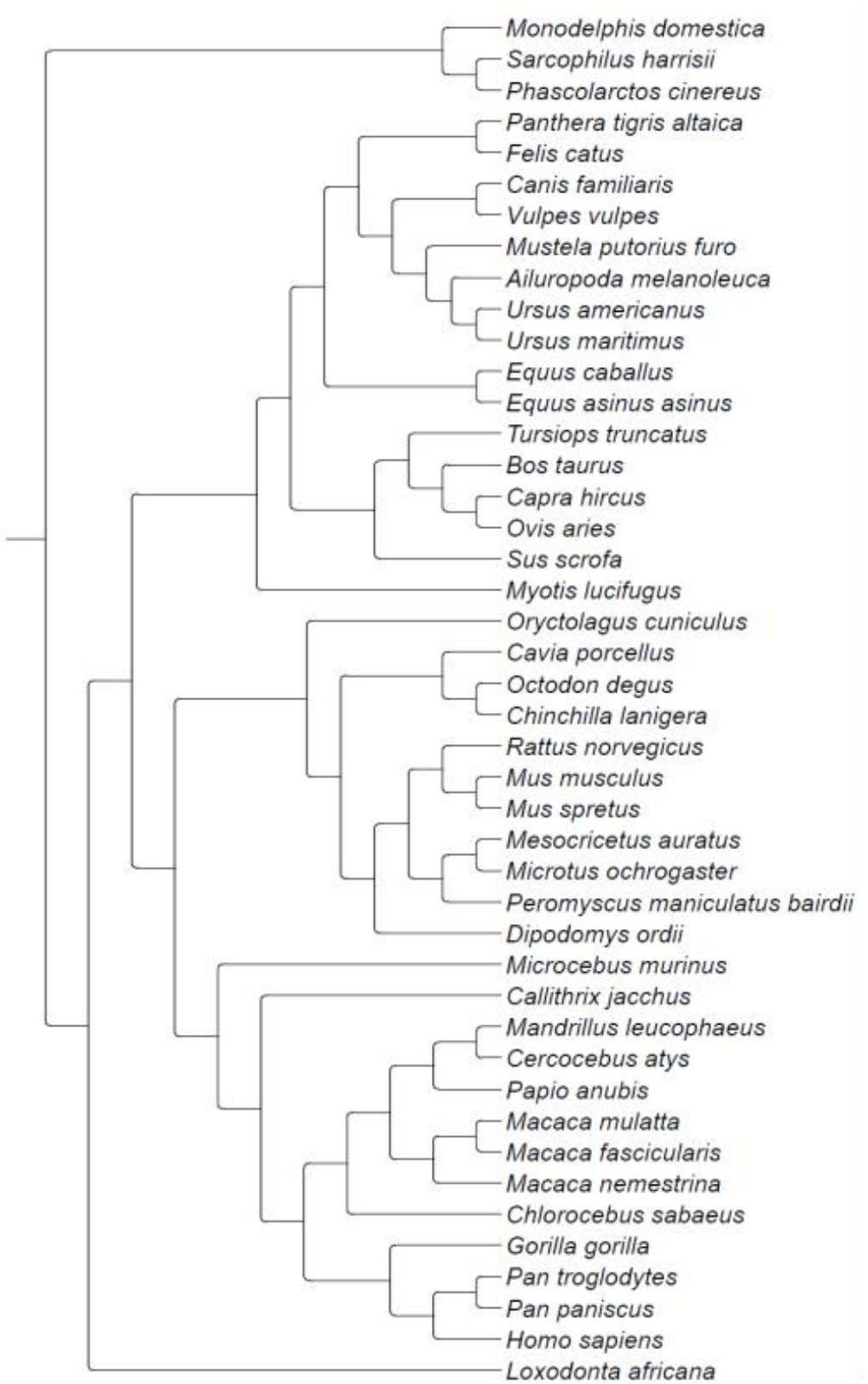


FIGURE S1. Phylogenetic tree of the set of species analysed. Distances obtained from the Timetree database (Kumar et al., 2017).

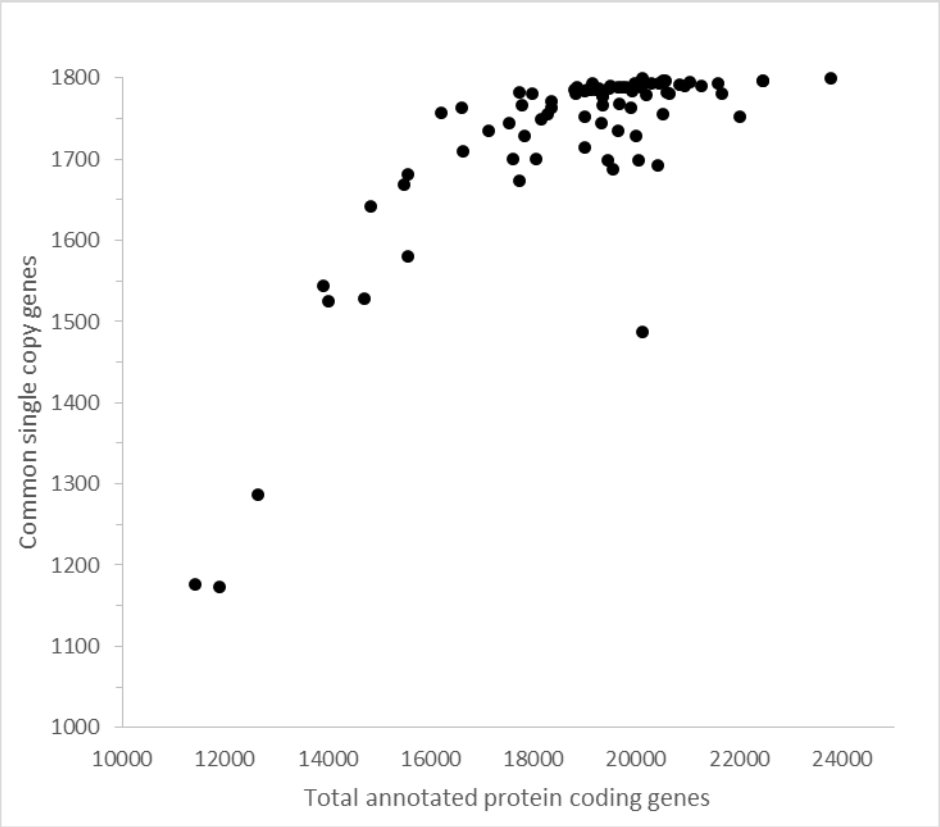


FIGURE S2. Plot of the total gene number of annotated genes per mammalian genome compared to the number of single copy core genes.

Chapter 4. Heritable transcriptome signatures of source population climatic conditions in lab-grown salt tolerant *Cakile maritima*

Abstract

Understanding the role of gene expression variation in plant adaptation is a major question in evolutionary biology. Transcriptional signatures associated with environment variation can derive from phenotypic plasticity and or heritable regulatory changes. In plants, the two effects can be separated by growing specimens under uniform conditions from seeds collected in different locations. *Cakile maritima* (sea rocket) is a succulent annual herb with a wide range coastal distribution and salt tolerance. To investigate gene expression variation in line with geographical distribution and climatic variables, here we analysed RNA-seq transcriptome profiles from 19 individuals of *C. maritima* derived from five different geographical regions. We found a significant association of gene expression variations in line with annual mean temperature, annual precipitation, mean diurnal range, temperature seasonality, mean temperature of wettest quarter and precipitation seasonality. Associated genes were significantly enriched in various functional categories related to plant stress responses, e.g. negative regulation of endopeptidase activity, polysaccharide catabolic process, glycolytic process, fucose metabolic process, cell redox homeostasis, regulation of gene expression and DNA duplex unwinding. Overall, these results provide novel insights into the genetic mechanisms of adaptation to changing environments in plants.

Key words: brassica, climate, local adaptation, dispersal, gene expression

Main points

- In this study we carry out transcriptome profile analyses from plants grown in controlled conditions from seeds collected from multiple locations around the world.
- Gene expression profiles of all *Cakile maritima* samples were found to be more similar to each other than to transcriptomes from related species.
- A set of genes had expression patterns which correlated with various aspects of local climate factors.
- This is the first assessment of transcriptome signatures of local adaptation in *Cakile maritima*, a salt and drought tolerant relative of model organism *Arabidopsis thaliana* and of other commercially important brassica crops.

Introduction

Plants are particularly affected by changes in their surroundings due to their sessile condition. Genetic variability and phenotypic plasticity drive the ability of plants species (Koenig and Weigel, 2015) to adapt to local conditions and thrive across broad geographic ranges (Halbritter et al., 2018, Leimu and Fischer, 2008) and determine whether populations persist under changing conditions, including climate change (Nunney, 2015, Hoffmann and Sgrò, 2011). For instance, several studies on genetic adaptation along elevation gradients indicate that this is a widespread phenomenon in plants as reviewed by Halbritter (2018). This ability for local adaptation depends on various factors including population size, gene flow, heritable/genetic variation, and the rate of environmental change (Williams et al., 2008, Leimu and Fischer, 2008).

Plant species have developed a wide range of mechanisms to respond to changing environments (Sham et al., 2015, Shao et al., 2007). Recently, Lobréaux and Miquel (2019b) revealed a link between differences in elevation and sequence variation in genes related to stress response in the plant *Arabis alpina* (Brassicaceae family) through a genome wide scan analysis. (Lobréaux and Miquel, 2019b). Similarly, a study comparing genomes from *Arabidopsis lyrata* plants revealed specific polymorphisms associated growing on serpentine and granitic soils (Turner et al., 2010). These polymorphisms were near and or within genes enriched in metal ion transmembrane transporter activity and calcium ion binding gene ontology terms.

Regulation of gene expression is considered an essential strategy for the adaptation to environmental stress conditions (Sham et al., 2015). Transcriptome analysis on diverse plant populations covering varied geographical and environmental gradients are emerging as a useful approach to enhance the understanding of environmental factors and genetic

mechanisms underlying locally adaptive trait variation (Akman et al., 2016). Analyses of mRNA-seq data from 19 phenotypic and ecologically diverse populations of the species *Protea repens* found an association between source population climate and gene expression patterns when plants were grown in controlled conditions. Gene expression patterns reflected heritable trait variation and population differentiation along environmental gradients (Akman et al., 2016). A separate study focused on the molecular adaptive mechanisms of desert plants by analysing gene expression profiles of the shrub *Artemisia sphaerocephala* in response to different stresses revealing a set of transcripts involved in heat, cold, salt and drought responses shedding light into the molecular mechanisms of adaptation of this desert plant (Zhang et al., 2016).

Cakile maritima (searocket) is a glabrous, succulent annual herb with a short life cycle (three months) and is of particular interest due to its coastal distribution and salt tolerance (Gandour et al., 2008, Clausing et al., 2000). *C. maritima* is able to migrate at an average rate of 53 km per year with its seeds surviving up to four months immersed or up to one year floating in sea water (Barbour and Rodman, 1970, Gandour et al., 2008). *C. maritima* can be found in a wide range of latitudes and shows considerable phenotypic diversity (Davy et al., 2006). It has been suggested that ocean currents prevents seed exchange between populations of the northern coast and the others resulting in geographic population structure (Ben Hamed et al., 2016). These characteristics make *C. maritima* an excellent system to investigate local adaptation to environment but the molecular mechanisms associated with this species dispersal and local adaptation have yet to be explored.

In the present study, we analyse transcriptome profiles of 19 *C. maritima* plants grown in controlled conditions from seeds collected in natural populations comprising varied geographical distributions. Using a phylogenetically corrected correlation approach, we test whether changes in

gene expression are associated with bioclimatic variation. The results improve our understanding in the molecular signatures of trait variation in *C. maritima* which have facilitated its adaptation to novel environments as this plant has been dispersed around the globe.

Hypotheses

1. If hereditary adaptations to local environments have taken place as *Cakile maritima* has dispersed around the world then we expect to observe clustering of transcriptome profiles according to location of origin of seeds collected for tissue sampling. We would also expect to observe that a proportion of genes would show an association between environmental factors and their expression levels.
2. If local adaptation has taken place which is related to local environment, then we would expect to observe a set of genes which are related to climatic variables even after taking into account phylogenetic interrelatedness between populations given the patterns of dispersal across the world.
3. If local adaptation related to distinct climatic variables then we expect specific sets of functional categories to be enriched among genes associated with each climatic variable.

Material and Methods

Illumina RNA sequencing

We used transcriptome data obtained from flower buds taken from 19 *C. maritima* and three other closely related *Cakile* species plants (*C. edentula*, *C. lanceolata* and *C. arabica*). All plants were grown together in a

greenhouse under uniform conditions from seeds collected on different geographical regions (Table 1).

Transcriptome data annotation

Read quality analysis was assessed on the raw RNA data using FastQC (version 0.11.4). Quality and adapter trimming was performed using Trim Galore (version 0.4.1), trimming low-quality ends (Phred score < 20), clipping adapters and reads shorter than 70 bp. Resulting unpaired reads were discarded. FastQC was used to re-evaluate the integrity of the trimmed reads prior to subsequent mapping and analysis. As the genome of *Cakile* species including *C. maritima* are not available, Illumina read data was annotated using the genome of the closely species *Brassica oleracea* (v2.1) (Kersey et al., 2015), retrieved from EnsemblPlants Biomart release 31 (Kinsella et al., 2011). Gene annotations from *B. oleracea* were also retrieved from the same source. Curated reads were mapped to the *B. oleracea* genome using the high accuracy and sensitive mapping method Stampy (version 1.0.28) (Lunter and Goodson, 2011) which is recommended for NGS data and annotation to a reference genome from a different species (Nielsen et al., 2011, Benjamin et al., 2014). Mapped read alignment SAM files of each sample were converted to BAM format with SAMtools toolkit on useGalaxy server (Afgan et al., 2016) (v 2.1), in addition to sorting, and filtering of unmapped reads or duplicates (MAPQ quality score <10) (v 1.1.2). The number of reads mapped to each gene was counted and reported using the HTseq package (v 0.6.0) on useGalaxy server, specifically the union mode of htseq-count was selected.

In order to filter low expressed genes we removed genes with zero variance and with less than 10 counts across the *C. maritima* samples. For this, first data was normalised to transcripts per kilobase million (TPM) values to account for between sample differences in sequencing depth and gene length. The removal of very low read counts minimizes the background noise

given that they cannot be reliably distinguished (McIntyre et al., 2011), and also reduces the number of statistical tests that need to be carried out in downstream analyses (Law et al., 2016). The number of genes used for analyses was reduced from 49,151 to 33,191. After gene filtering, data was reverted to original read count number.

Transcriptome profile clustering

In order to explore the similarity among the samples we carried out a Principal Component Analysis (PCA) on read counts after *regularized logarithm* transformation (rlog) was performed using the DESeq2 package (Love et al., 2014). The transformation is recommended for clustering and ordination methods in RNA-seq data (Love et al., 2015).

Environmental data

Bioclimatic variables were compiled from the Worldclim Global Climate Database (http://www.worldclim.org) (Fick and Hijmans, 2017) with a spatial resolution of 5 minutes (of a longitude/latitude degree) (about 9 km at the equator) for each *C. maritima* plant sample based on the geographical coordinates of seed collection. The data represents the average for the years 1970-2000. Covariance among the 19 bioclimatic variables was tested by calculating a Pearson correlation coefficients matrix (Supplementary table 1). In order to minimize the multicollinearity among climatic variables that can affect correlation analysis, we removed highly correlated variables with a coefficient $|r| > 0.70$ as recommended in (Dormann et al., 2013b). Six bioclimatic variables were selected: annual mean temperature, mean diurnal range, temperature seasonality, mean temperature of wettest quarter, annual precipitation and precipitation seasonality (Supplementary Table 2). These variables represent annual trends (e.g., mean annual temperature, annual precipitation), seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of wettest quarter).

Phylogenetic controlled correlations of gene expression and bioclimatic variables

To account for phylogenetic relatedness influencing the degree of association between different parameters, phylogenetic independent contrasts (PIC) (Felsenstein, 1985) correction was implemented using the package "ape" (Paradis and Schliep, 2018) in R (CoreTeam, 2013). This method is widely used for assessing relatedness between variables while accounting for shared phylogenetic paths (Felsenstein, 1985). Genes were deemed to be associated with a specific bioclimatic variable based on a static threshold of a large effect size as defined by Cohen (r > 0.5) (Cohen, 1988). Focusing on effect sizes, instead of on p values reduces the probability of type two errors, where an alternative hypothesis would be wrongly rejected, particularly when sample sizes are low as is the case in this study (n = 19) (Nakagawa, 2004). To calculate the number of samples required to achieve the recommended statistical power of 0.8 (Cohen, 1988) when testing associations for 33191 genes, we carried out a power analysis using the following formula from (Cummings and Hulley, 1988a) as implemented in (http://www.sample-size.net/correlation-sample-size/):

$$N = \left[\frac{Z_\alpha + Z_\beta}{C}\right]^2 + 3$$

Where Z_α is the standard normal deviate for the significance threshold 0.05 divided by the number of tests carried out; Z_β is the normal deviate of the accepted level of type two errors (0.2 for a statistical power of 0.8) and C is calculated as follows:

$$C = 0.5 * \ln\left[\frac{1+r}{1-r}\right]$$

Where r refers to the size of associations for which significance should be reliably established.

Data for over 109 samples would be needed to achieve the recommended statistical power of 0.8 (Cohen, 1988) in identifying large (r > 0.5) when testing associations for a total of 33,191 genes.

To account for the possible high collinearity among gene expression profiles of non-independently varying genes, we calculated the effective number of tests based on the eigenvalues calculated from a correlation matrix of expression profiles per gene against expression profiles of all other genes (Li and Ji, 2005). The effective number of tests remains high at 29,846 and leaving the minimum number of samples needed to achieve adequate statistical power to detect large effect size associations (r > 0.5) largely unchanged at 108.

Gene ontology term enrichment analysis

Gene ontology (GO) annotations for each *B. oleracea* gene were obtained from the EnsemblPlants Biomart database release 43 (Kinsella et al., 2011). To avoid testing overrepresentation of GO functions associated with only a few annotated genes, GO terms associated with fewer than 50 associated gene genes were pooled together into a single category labelled “small GO” as applied in Castillo-Morales *et al*. (2014). Unlike the case in other studies, genes not associated with any functional GO term were included in the analyses under an “unknown GO” term. Enrichment analysis of GO categories among the set of genes was carried out by counting the number of genes assigned to each GO term within the analysed set of genes by contrasting to the proportion of genes associated to each GO term in 1,000 randomly selected samples of the same size. The mean and standard deviation of GO term representation in random generated datasets were taken to determine the corresponding p-values for each GO term using Z-score and Benjamini-Hochberg correction to correct for multiple testing as implemented in Castillo-Morales *et al*. (2014).

Variant calling

The Genome Analysis Toolkit (GATK, version 3.7) was used for calling RNA variants. Following the GATK Best Practices recommendation, post-alignment processing was implemented by marking duplicates and sorting the aligned reads with Picard Tools. Next, the SplitNCigarReads tool was used to split reads into exon segments and clip sequences, which overhang intronic regions. Variants were called implementing the Haplotype Caller in tool as has been shown to have the best performance compared to other variant callers (Liu et al., 2013, Pirooznia et al., 2014).

Phylogenetic tree construction

The variant calling files (gVCF) were combined and then filtered as follows: for positions where there a read in at least one sample, for positions where all samples have calls, for position where all calls were supported by more than five reads. These were then passed through SNPphylo that uses a maximum likelihood method for the inference of phylogeny (Lee et al., 2014).

Results

In order to assess the heritable transcriptional signature associated with adaptation to local climatic conditions in *C. maritima* , illumina RNA-seq data was obtained from flower buds from plants grown under uniform lab conditions from seeds collected in the wild from different geographical areas (Table 1). Short read transcriptome data was annotated against *Brassica oleracea,* the closest species with a sequenced genome available, identifying a total of 49,151 genes.

A principal component analysis (PCA) reveals all 19 *C. maritima* samples to be distinct from three closely related *Cakile* species, *C. arabica*, *C. edentula* and *C. lanceolata,* grown at the same time and under the same controlled conditions (Figure 1). A phylogenetic tree constructed from RNA read variants called shows *C. maritima* form four distinct clusters in agreement with their geographic origin (Figure 2). The South America, Australia, Mediterranean samples form three clear clusters with the UK and the Baltic sample forming a fourth group. The non-*C. maritima* samples fall outside all *C. maritima* clusters consistent with the fact of being separate species.

To evaluate the relationship between gene expression patterns and bioclimatic variables, correlation coefficients between gene expression and bioclimatic parameters were estimated for each sample. Variables examined were annual mean temperature, mean diurnal range, temperature seasonality, mean temperature of wettest quarter, annual precipitation and precipitation seasonality (Supplementary Table 2). Since the associations between variations in gene expression and the variables studied could be partly dependent on phylogenetic relatedness, we used a phylogenetic independent contrasts (PIC) test (CoreTeam, 2013) to account for this effect.

Genes associated with each variable were identified using a large effect size threshold ($r > 0.5$). A total of 56, 126 and 402 genes were found to be positively associated to annual mean temperature, mean diurnal range and temperature seasonality, respectively. Mean temperature of wettest quarter was found to be positively associated with 146 genes while annual precipitation to 269 and precipitation seasonality to 145 (Figure 3). Functional characterization of genes associated with each bioclimatic variable identified several functional gene ontology (GO) categories to be enriched (Figure 4). These include several terms related to plant responses to stress, like DNA duplex unwinding, sugars metabolic processes (glucose,

fucose and polysaccharide), cell redox homeostasis, regulation of endopeptidase activity and regulation of gene expression.

Discussion

The present work investigated the transcriptional architecture of the searocket (*Cakile maritima*) and the relationship with its wide geographical distribution by studying the gene expression of over 49,000 genes of 19 different samples from five different regions. The study of the relationship between genetic and environmental variation is fundamental to comprehending the mechanisms underlying plant adaptation. The increasing use of transcriptomic data has allowed the discovery of molecular mechanisms that enhance our knowledge of trait variation and adaptation (Akman et al., 2016)(Akman et al., 2016)(Akman et al., 2016)(Schoville et al., 2012, Pespeni et al., 2013, Yang et al., 2015, Akman et al., 2016).

Phylogenetic reconstruction reveals a pattern of genetic divergence between geographical areas with four evident groups: South America, Australia, Mediterranean and the UK samples forming a cluster with the Baltic sample. Samples from three different species fall outside all *C. maritima* clusters. Transcriptome profile analysis shows a consistent differentiation of *C. maritima* from samples from other species but geographical groupings are less apparent. This result means that the first expectation under hypothesis one is not met. Although there is a set of genes with strong effect size associations with various climatic variables after controlling for the phylogenetic interrelatedness among individuals sampled. This is consistent with expectations under hypothesis two. Previous studies of transcriptome and genetic variation across populations of a single species have shown that gene expression profiles have a less apparent clustering than genetic variation analyses [e.g. (Martin et al., 2014)].

Genes associated with each bioclimatic variable are significantly enriched in various functional categories previously linked to differences in environmental conditions in several plant species (Raikwar et al., 2015, Koiwa et al., 2003, Wan et al., 2018, Keunen et al., 2013, Hieng et al., 2004). This is consistent with expectations under hypothesis three. For instance, among the genes associated to temperature seasonality, there is enrichment in biological processes related to plant stress responses (negative regulation of endopeptidase activity, polysaccharide catabolic process, cell redox homeostasis, tRNA aminoacylation for protein translation and regulation of gene expression). Proteolytic activity has been described as an essential mechanism for plant adaptation to different environmental conditions given that protein breakdown plays an important role in cellular housekeeping by the removal of unwanted proteins, signal propagation, reutilization of amino acids, and modification of protein content for required changes in metabolic status during certain conditions (Vierstra, 1996, Hieng et al., 2004). Moreover, the activity of peptidases can be up or down regulated depending on the plant sensitivity to stress (Hieng et al., 2004).

Likewise, the maintenance of the redox environment in plant cells is essential to overcome stressful environmental conditions (Keunen et al., 2013). Exposure to oxidative cellular environments including extreme temperature generates toxic reactive oxygen species (ROS) (Suzuki and Mittler, 2006). The role of these molecules can be detrimental by oxidative injure of cells or beneficial by acting as main signalling regulators of defence pathways leading to cellular protection and/or acclimation (Keunen et al., 2013, Miller et al., 2008). Interestingly, it has been observed that plant sugars can contribute to both activities (Keunen et al., 2013). Remarkably, in addition to temperature seasonality, mean diurnal range and mean temperature of wettest quarter also show a significant enrichment of associated genes in processes involved in sugar metabolism (glycolytic process and fucose metabolic process). These findings highlight the importance of modulating regulatory mechanisms and signal transduction pathways to maintain a

normal physiological steady state in changing environmental conditions affected by temperature increase. Glycolitic process related gene PFP-ALPHA1 (Bo8g070530) that codes for the protein pyrophosphate: fructose 6-phosphate 1-phosphotransferase (PFP1) was found to be associated with mean diurnal range. Evidence from *in vivo* experiments in *Daucus carota* shows that PFP1 acts as important sensor of environmental changes and helps the mobilisation of energy reserves during unfavourable environments (Lim et al., 2009).

Genes associated to precipitation seasonality are enriched in functional annotations related to proteolytic activity (negative regulation of endopeptidase activity and cellular amino acid metabolic process) and regulation of ROS (hydrogen peroxide catabolic process). Proteolysis has been previously linked to management of water availability in plants. For instance, experiments on *Phaseolus vulgaris* with differing sensitivities to water withdrawal revealed that serine proteinase activity is involved in the response to drought stress (Hieng et al., 2004). Given that *C. maritima* inhabits water scarce ecosystems, it seems likely that proteolysis and ROS regulation are important biological processes in the adaptation to intra-annual precipitation instability.

Regulation of gene expression is an essential strategy for the adaptation to environmental stress conditions influenced by biotic and abiotic stress factors (Sham et al., 2015, Akman et al., 2016). This modulation involves the transcriptional, post-transcriptional and post-translational levels (Lopez-Maury et al., 2008). Cold acclimation, the process of acquiring freezing tolerance after being exposed to low non-freezing temperatures, has been associated with changes in gene expression. For example, research on *Arabidopsis thaliana* showed that *frostbite 1* (*fro1*) mutant plants subject to low-temperature treatment presented constitutive accumulation of ROS, and reduced expression of cold-responsive genes (including *RD29A*, *COR47*,

COR15A, and *KIN1*) thus affecting its capacity for cold acclimation (Lee et al., 2002). Regulation of gene expression in response to low temperature stress has also been observed at the post-transcriptional level, particularly by microRNAs through mRNA cleavage and translational repression (Megha et al., 2018).

Other important regulators of stress-induced pathways in plants are helicases. These enzymes catalyze the unwinding of energetically stable duplex DNA and are essential for basic cellular processes regulating plant growth and development (Tuteja et al., 2012). The functional category DNA duplex unwinding is enriched among genes associated to mean temperature of wettest quarter. This bioclimatic variable provides a measure of the mean temperatures that prevail during the wettest season, therefore the observed relationship could imply a role of DNA unwinding in response to extreme or limiting environmental factors.

The results of this study show that phenotypic variation in response to local conditions is at least partially heritable. Tissue samples collected from specimens grown in common and controlled conditions from seeds collected from different environments shows a transcriptional signal of source population climate. This is consistent with population differentiation and adaptation to local conditions through heritable changes in gene expression. This is true even after correcting for phylogenetic similarity across populations. Moreover, enrichment of genes associated with bioclimatic variables were found to be associated with various aspects of plant stress responses suggesting a role of transcriptional profile changes in adaptation of *C. maritima* to local environmental conditions. The findings support the importance given to high temperatures and water availability as part of the key environmental stresses affecting *C. maritima* populations that occupy fragile habitats in dunes and beaches (Ciccarelli et al., 2010). This could have important implications for effective coastal habitat management given

the role of this plants in the building and stabilization of dune ecosystems and their particular sensitivity to climate change (Frosini et al., 2012).

Given that the data derives from flower bud samples, the results from these analyses reveal a specific part of the differentiated gene expression in multiple environments. It would be interesting to conduct the same approach on other tissues to complement the study. Importantly, genes identified as associated with bioclimatic variables could prove valuable in understanding the response to drought and salt tolerance in economically important crops as *C. maritima* along with model plant *A. thaliana* is a member of the family Brassicaceae (Cruciferae) which also includes broccoli, cauliflower, cabbage, choy sum, rutabaga, turnip and some seeds used in producing canola oil.

Altogether, these results provide novel insights into the molecular basis of adaptation to changing environments in plants. Furthermore, identifying genomic signatures associated with the adaptation to varying environmental conditions critically contributes to our understanding of the vulnerability of plant species to climate change.

Tables

TABLE 1. Geographical coordinates of the *Cakile maritima* samples analysed.

Region	Sample	Total number of reads	Latitude	Longitude
Australia	AUS1	18,338,082	35.8811° S	150.1541°E
	AUS2	20,606,034	33.8060° S	151.2948°E
	AUS3	17,543,721	35.5245° S	150.4028°E
	AUS4	14,006,510	38.3726° S	178.2968°E
South America	SA1	13,807,105	32.2113° S	52.1819°W
	SA2	6,266,356	33.5216° S	53.3665°W
	SA3	14,446,538	38.0055° S	57.5426°W
	SA4	12,000,805	32.2113° S	52.1819°W
Mediterranean	MED1	18,498,599	38.2498° N	0.5157°W
	MED2	17,138,916	39.7538° N	19.6416°E
	MED4	17,062,583	35.9375° N	14.3754°E
	MED5	17,901,014	39.2979° N	9.6094°E
	MED6	13,422,280	30.6085° N	33.6176°E
United Kingdom	UK1	6,490,321	55.6377° N	4.7835°W
	UK2	5,513,453	53.1464° N	0.3379°E
	UK3	16,913,035	51.1174° N	4.2056°W
	UK4	15,615,821	50.9613° N	0.9624°E
	UK5	18,871,595	51.5217° N	3.7413°W
Baltic	BAL	19,327,293	59.4370° N	24.7536°E

Figure legends

FIGURE 1. PCA analysis of *Cakile* normalized gene counts for all samples. Two-dimensional principal component analysis (PCA) results based on PC1 and PC2. Numbers in parentheses represent the percentage of total variance explained by each principal component (PC). *Cakile maritima* [Australia (AUS); South America (SA); Mediterranean (MED); United Kingdom (UK); Baltic (BAL)]; *Cakile lanceolata* (LAN); *Cakile arabica* (ARA); *Cakile edentula* (EDE).

FIGURE 2. Phylogenetic tree of diverse *Cakile* samples obtained by maximum likelihood inference with SNPhylo. *Cakile maritima* [Australia (AUS); South America (SA); Mediterranean (MED); United Kingdom (UK); Baltic (BAL)]; *Cakile lanceolata* (LAN); *Cakile arabica* (ARA); *Cakile edentula* (EDE).

FIGURE 3. Relationship between changes in gene expression and bioclimatic variables in *Cakile maritima*. Venn diagram comparing genes displaying a positive association between expression and each of the different bioclimatic parameters. Genes associated with annual mean temperature (n = 56) and annual precipitation (n = 269) (not shown in figure) do not overlap with the gene sets associated with other variables.

FIGURE 4. Gene ontology categories of the genes with expression levels associated with bioclimatic variables. Heat map showing over-representation of biological process GO terms among genes displaying a significant positive association between transcript abundance (counts) and bioclimatic variables ($p < 0.05$).

Figures

FIGURE 1.

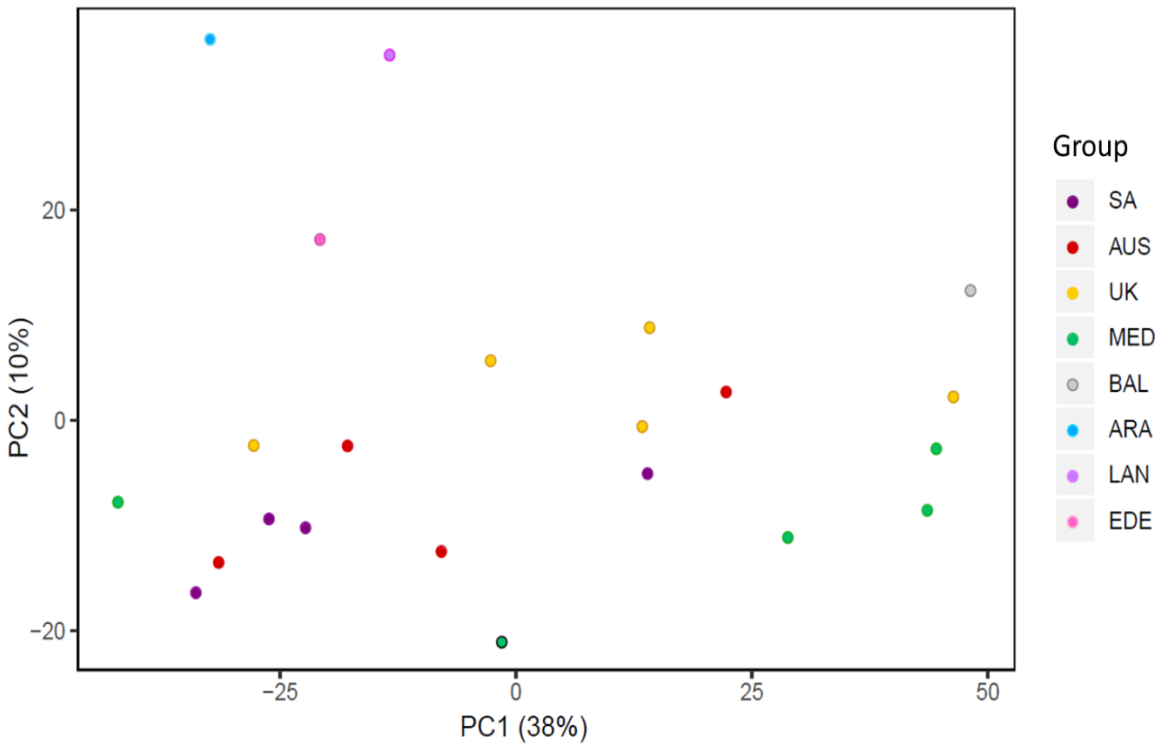


FIGURE 2.

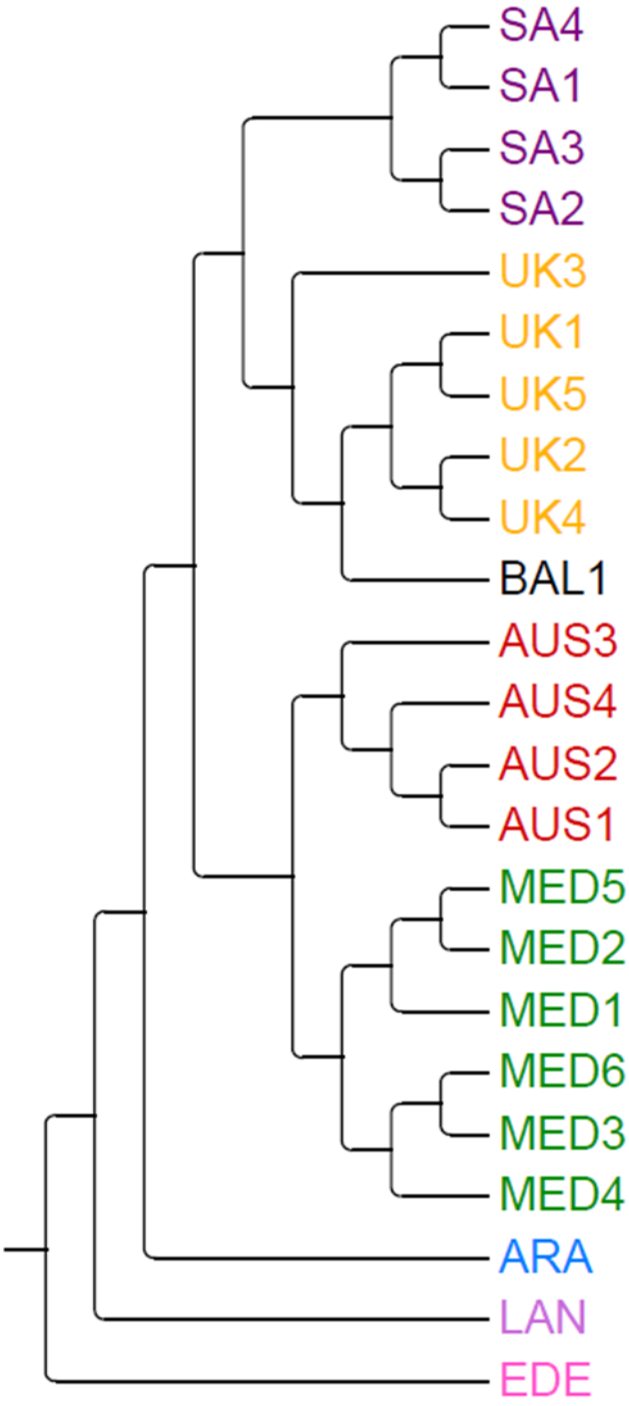

SA4
SA1
SA3
SA2
UK3
UK1
UK5
UK2
UK4
BAL1
AUS3
AUS4
AUS2
AUS1
MED5
MED2
MED1
MED6
MED3
MED4
ARA
LAN
EDE

FIGURE 3.

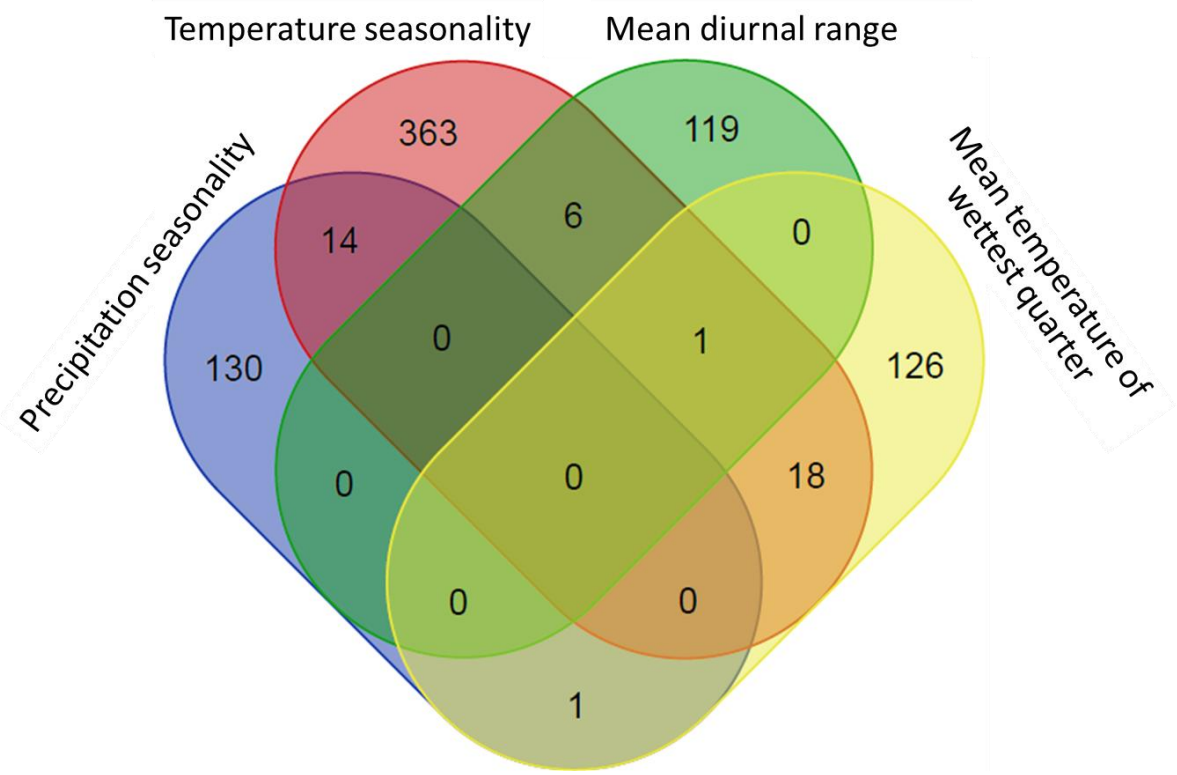
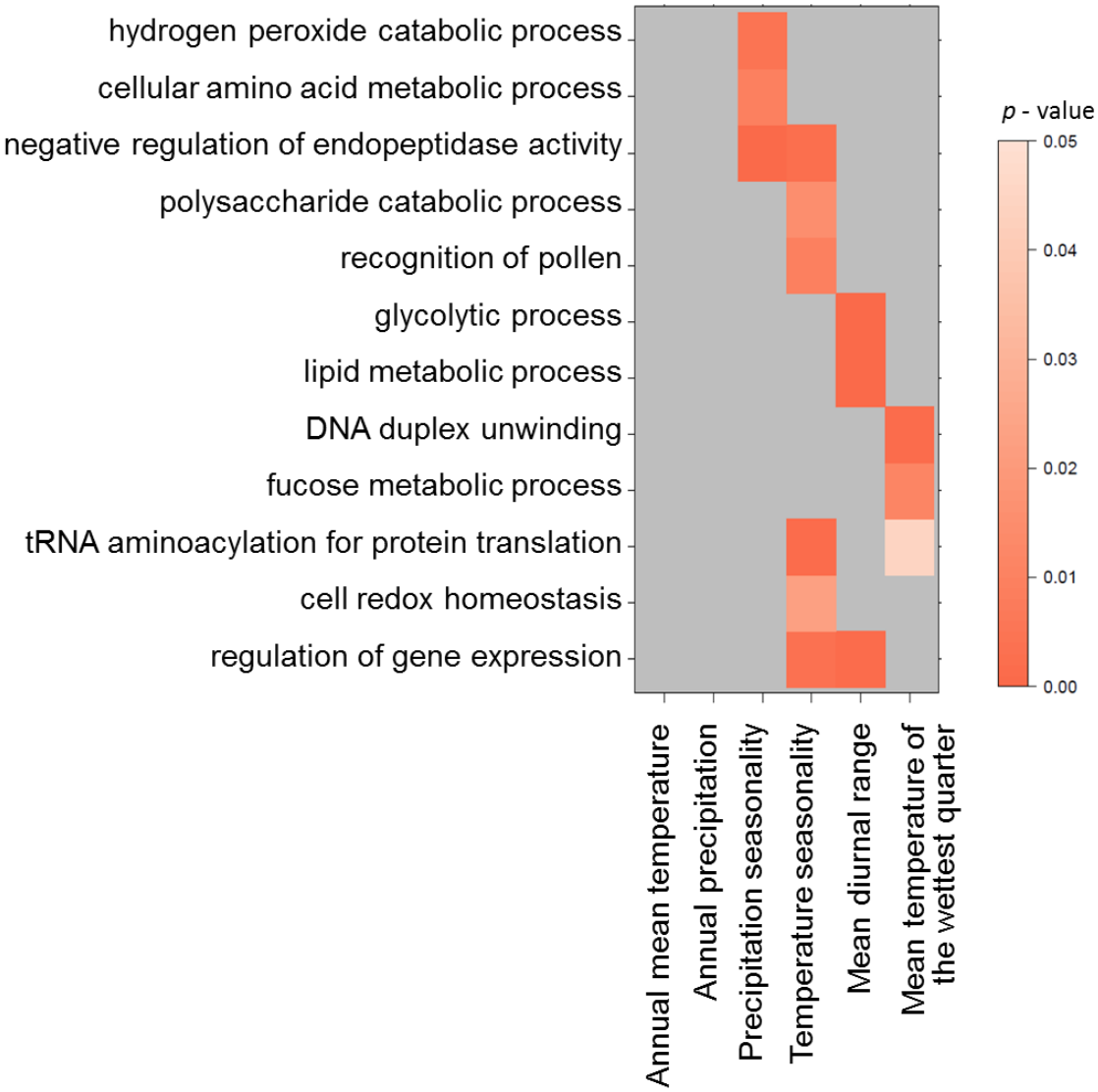

Temperature seasonality
Mean diurnal range
Precipitation seasonality
Mean temperature of wettest quarter
363
119
14
6
0
130
0
1
126
0
0
18
0
0
1

FIGURE 4.


hydrogen peroxide catabolic process
cellular amino acid metabolic process
negative regulation of endopeptidase activity
polysaccharide catabolic process
recognition of pollen
glycolytic process
lipid metabolic process
DNA duplex unwinding
fucose metabolic process
tRNA aminoacylation for protein translation
cell redox homeostasis
regulation of gene expression
Annual mean temperature
Annual precipitation
Precipitation seasonality
Temperature seasonality
Mean diurnal range
Mean temperature of the wettest quarter
p - value
0.05
0.04
0.03
0.02
0.01
0.00

Supplementary Material

TABLE S1. Bioclimatic variables for the *Cakile maritima* samples analysed.

Sample	BIO1	BIO2	BIO3	BIO4	BIO5	BIO6	BIO7	BIO8	BIO9	BIO10	BIO11	BIO12	BIO13	BIO14	BIO15	BIO16	BIO17	BIO18	BIO19
SA1	181	81	42	3727	280	90	190	138	214	229	133	1225	126	66	17	353	236	311	344
SA2	169	91	46	3488	277	82	195	129	156	209	125	1180	116	89	8	314	272	298	311
SA3	134	103	46	4176	255	33	222	171	82	187	82	873	98	54	17	260	172	252	172
SA4	181	81	42	3727	280	90	190	138	214	229	133	1225	126	66	17	353	236	311	344
AUS1	158	90	48	3410	245	60	185	189	118	201	113	1077	128	56	24	340	178	333	228
AUS2	175	92	47	3703	262	69	193	218	132	219	125	1294	155	65	27	429	225	356	305
AUS3	160	88	48	3413	247	64	183	190	120	202	115	1217	136	68	23	379	209	361	275
AUS4	142	98	50	3372	245	49	196	99	156	185	99	1430	177	75	30	513	244	259	513
UK1	83	67	36	4284	181	-1	182	59	104	140	30	1334	147	72	26	432	225	279	371
UK2	97	69	33	4815	208	2	206	45	61	159	36	606	61	38	13	171	127	157	148
UK3	109	58	33	4343	204	29	175	65	124	165	55	1038	120	60	25	352	183	210	312
UK4	104	70	35	4791	214	15	199	80	121	166	45	667	78	43	22	218	133	140	177
UK5	99	61	32	4506	202	17	185	54	118	159	44	1171	135	69	24	392	212	238	349
MED1	181	107	41	5230	317	60	257	195	248	251	117	320	61	4	54	138	28	48	75
MED2	165	88	35	5510	303	56	247	114	238	238	100	1133	191	12	65	531	52	52	472
MED4	185	72	34	4982	303	93	210	171	248	252	126	514	96	0	82	269	8	41	171
MED5	174	70	33	5045	294	83	211	155	239	243	116	308	46	2	53	124	16	46	91
MED6	196	104	42	5053	319	73	246	128	254	256	128	67	13	0	88	37	0	0	37
BAL	53	66	22	7989	211	-86	297	147	-15	158	-46	672	80	32	31	230	101	209	137

NOTE: Annual Mean Temperature (BIO1), Mean Diurnal Range (Mean of monthly (max temp - min temp)) (BIO2), Isothermality (BIO2/BIO7) (* 100) (BIO3), Temperature Seasonality (standard deviation *100) (BIO4), Max Temperature of Warmest Month (BIO5), Min Temperature of Coldest Month (BIO6), Temperature Annual Range (BIO5-BIO6) (BIO7), Mean Temperature of Wettest Quarter (BIO8), Mean Temperature of Driest Quarter (BIO9), Mean Temperature of Warmest Quarter (BIO10), Mean Temperature of Coldest Quarter (BIO11), Annual Precipitation (BIO12), Precipitation of Wettest Month (BIO13), Precipitation of Driest Month (BIO14), Precipitation Seasonality (Coefficient of Variation) (BIO15), Precipitation of Wettest Quarter (BIO16), Precipitation of Driest Quarter (BIO17), Precipitation of Warmest Quarter (BIO18), Precipitation of Coldest Quarter (BIO19). Temperature data in °C * 10. Precipitation data in mm (millimetre).

TABLE S2. Pearson's correlation coefficients between bioclimatic variables for the *Cakile maritima* samples analysed.

AMT	1																		
MDR	0.62	1																	
ISO	0.59	**0.76**	1																
TS	-0.43	-0.29	**-0.82**	1															
MxTwaM	**0.90**	0.64	0.32	-0.02	1														
MnTcoM	**0.94**	0.44	0.62	-0.64	**0.73**	1													
TAR	-0.09	0.25	-0.42	**0.85**	0.33	-0.40	1												
MTwtQ	0.60	0.59	0.45	-0.08	0.60	0.42	0.23	1											
MTdrQ	**0.84**	0.36	0.24	-0.20	**0.83**	**0.82**	-0.01	0.25	1										
MTwaQ	**0.93**	0.55	0.31	-0.06	**0.98**	**0.78**	0.24	0.63	**0.87**	1									
MTcoQ	**0.97**	0.60	**0.71**	-0.63	**0.78**	**0.98**	-0.30	0.53	**0.79**	**0.82**	1								
AP	-0.14	-0.08	0.38	-0.56	-0.40	0.03	-0.58	-0.12	-0.28	-0.38	0.02	1							
PwtM	-0.05	-0.02	0.31	-0.42	-0.23	0.07	-0.41	-0.05	-0.10	-0.21	0.08	**0.91**	1						
PdrM	-0.26	-0.09	0.42	-0.62	-0.55	-0.07	-0.65	-0.21	-0.47	-0.55	-0.07	**0.86**	0.59	1					
PS	0.43	0.21	-0.19	0.40	0.63	0.26	0.48	0.22	0.64	0.64	0.27	-0.63	-0.33	**-0.85**	1				
PwtQ	-0.10	-0.08	0.27	-0.40	-0.28	0.04	-0.43	-0.11	-0.13	-0.26	0.03	**0.92**	**1.00**	0.61	-0.33	1			
PdrQ	-0.20	-0.07	0.44	-0.64	-0.50	-0.02	-0.65	-0.17	-0.42	-0.49	-0.02	**0.88**	0.62	**0.99**	**-0.85**	0.63	1		
PwaQ	-0.16	-0.01	0.47	-0.58	-0.46	-0.04	-0.56	0.12	-0.50	-0.42	0.00	**0.82**	0.57	**0.91**	**-0.81**	0.57	**0.92**	1	
PcoQ	-0.07	-0.08	0.26	-0.45	-0.24	0.10	-0.46	-0.29	-0.02	-0.24	0.07	**0.90**	**0.94**	0.63	-0.35	**0.95**	0.66	0.50	1
	AMT	MDR	ISO	TS	MxTwaM	MnTcoM	TAR	MTwtQ	MTdrQ	MTwaQ	MTcoQ	AP	PwtM	PdrM	PS	PwtQ	PdrQ	PwaQ	PcoQ

NOTE: Annual mean temperature (AMT), mean diurnal range (MDR), isothermality (ISO), temperature seasonality (TS), max temperature of warmest month (MxTwaM), min temperature of coldest month (MnTcoM), temperature annual range

(TAR), mean temperature of wettest quarter (MTwtQ), mean temperature of driest quarter (MTdrQ), mean temperature of warmest quarter (MTwaQ), mean temperature of coldest quarter (MTcoQ), annual precipitation (AP), precipitation of wettest month (PwtM), precipitation of driest month (PdrM), precipitation seasonality (PS), precipitation of wettest quarter (PwtQ), precipitation of driest quarter (PdrQ), precipitation of warmest quarter (PwaQ), precipitation of coldest quarter (PcoQ). Correlation coefficients with $|r| > 0.70$ are bold-faced. Variables selected are underlined.

TABLE S3. Bioclimatic variables for the *Cakile maritima* samples analysed.

Sample	AMT (°C)	MDR (°C)	TS (°C)	MTwtQ (°C)	AP (mm)	PS (%)
AUS1	158	90	3410	189	1077	24
AUS2	175	92	3703	218	1294	27
AUS3	160	88	3413	190	1217	23
AUS4	142	98	3372	99	1430	30
BAL1	53	66	7989	147	672	31
MED1	181	107	5230	195	320	54
MED2	165	88	5510	114	1133	65
MED4	185	72	4982	171	514	82
MED5	174	70	5045	155	308	53
MED6	196	104	5053	128	67	88
SA1	181	81	3727	138	1225	17
SA2	169	91	3488	129	1180	8
SA3	134	103	4176	171	873	17
SA4	181	81	3727	138	1225	17
UK1	83	67	4284	59	1334	26
UK2	97	69	4815	45	606	13
UK3	109	58	4343	65	1038	25
UK4	104	70	4791	80	667	22
UK5	99	61	4506	54	1171	24

NOTE: Annual Mean Temperature (AMT), Mean Diurnal Range (MDR), Temperature Seasonality (TS), Mean Temperature of the Wettest Quarter (MTwtQ), Annual Precipitation (AP), Precipitation Seasonality (PS). Temperature data in °C * 10. Precipitation data in mm (millimetre).

Chapter 5. General conclusions

Throughout the diverse studies carried out in the present work we have showed how distinct genomics approaches employing different types of data, *i.e.,* genomic and transcriptomic data, focusing in a single species or comparing several can help to elucidate the genetic basis of complex phenotypes that play important roles in the evolutionary history of eukaryotes. The analyses performed when examining three different phenotypes revealed links between differences in genetic features, both at sequence and expression levels, with changes in phenotypes and geographical distribution that might influence their adaptation and divergence.

In Chapters 2 and 3 of this thesis, I analysed available genomic mammalian data in order to increase our understanding of the genomic bases of two important complex features in the evolution of this taxa, brain evolution and sexual size dimorphism as indicator of sexual selection. I identified a set of genes associated with each of the two sets of phenotypes under study which were significantly overexpressed in cellular functions related with brain development and the immune response in the case of the brain related phenotypes. Enrichment analyses in association with sexual selection revealed that less sexually dimorphic species are found to have an expansion of gene families associated with brain development.

In the first study (Chapter 2), I addressed the marked changes in cellular composition of the brain in mammals by comparing differences at the genome level. Specifically, I examined gene family size (GFS) variations to investigate expansion signals of gene families over time and its probable relationship with brain function. Gene family repertoires, determined by gene gain and death rates, are variable across different lineages and may influence adaptation or speciation (Lynch and Conery, 2000). Gene family

expansion due to duplication or *de novo* formation of genes might lead to the creation of genes that help to wider biological functions, reflecting changes in the relative relevance of the molecular functions they represent. It is worth noting that because GFS analyses rely on evidence of gene presence and absence, then derived results are susceptible to bias due to quality of genome sequences and assemblies. Accordingly, we removed species missing a high number of “single copy core” as they likely represent low quality genomes rather than actual variations in gene number. Enrichment of gene families associated with various aspects of brain development was found in association with encephalization and cell composition parameters: neuron number, glia to neuron ratio and neuron density. The results also show excess of gene families associated with encephalization and neuron number compared to chance expectations being enriched in immune system processes. Expansion of gene families associated with the immune system in line with encephalization has been previously shown in a smaller set of species (Castillo-Morales et al., 2014, Castillo-Morales et al., 2016). Sequential regressions and general linear models including both phenotypes in combination with other cell composition parameters (neuron density and glia to neuron ratio) reveal that gene families associated with encephalization and neuron number are independently associated with these functions. Gene families associated with neuron density were enriched in cell projection and neuron development whereas families associated with glia to neuron ratio were enriched in translation and cell migration. Phylopath analysis results are consistent with a causal link between changes in gene number increases in encephalization associated gene families and encephalization but not driving the control variable body mass. As additional data on cell composition accumulates in the future, it will be possible to test causal associations in other phenotypes. Examining gene families associated with encephalization we found particular examples (BCL2, SHANK and MDB) of families related to brain development and linked to neurodegenerative processes as well as neurodevelopmental and neuropsychiatric disorders. It would be interesting to examine whether expanding gene families in line with brain evolution are disproportionately associated with psychiatric disorders and/or

neurodegeneration. Together, these results represent the first genomic screening for gene families associated with evolutionary variations in cellular composition of the brain. Importantly, the use of a comparative approach allows the identification of gene families and molecular functions specifically associated with the evolution of larger and more complex brains as is the human's which may not be uncovered from murine model focused research and which could be relevant to disease states.

Chapter 3 focused on the study of the molecular mechanisms that influence differences in sexual dimorphism (SSD), an indicator of sexual selection. Thus, implementing a comparative genomics approach in 44 mammalian species we observed evidence of gene family expansion coupled with variations in sexual size dimorphism. Interestingly, within the expanded gene families most strongly associated to decreased dimorphism, there is a significant overrepresentation of families associated with biological functions specifically related to brain development. No such associations were found for families with the greater expansion among species with the greatest dimorphism. Gene turnover rates and the evolution of gene family sizes are important aspects of genome evolution. Particularly, the results of this study are relevant to the study of gene family expansions related to the evolution of sexual selection and support the idea of a complex interplay in the evolution of this mode of natural selection with body mass and brain complexity.

The association between sexual size dimorphism (SSD) with species body mass has been shown in many distinct groups of mammals (Weckerly, 1998, Kappeler et al., 2019, Lindenfors et al., 2007). Across this group, larger species tend to also be the most dimorphic in body size, trend known as Rensch's rule (Rudoy and Ribera, 2017, Fairbairn, 1997, Lindenfors et al., 2007). Although several hypotheses have been proposed to explain variations in body mass among species, sexual selection has been proposed to be a fundamental driver of body size increase (Blanckenhorn, 2000).

Lower rates of body size dimorphism is primarily seen in species with monogamous mating systems and often, with bi-parental care for offspring (Weckerly, 1998, Pérez-Barbería et al., 2002, Kleiman, 1977). It is likely that such close social bonds require increasingly complex social skills leading to changes in brain circuitry.

Despite the recognized importance of sexual selection and brain evolution variation among species, few studies have approached the relationship between these two features (Schillaci, 2006, Pitnick et al., 2006, Garamszegi et al., 2005, Madden, 2001). Research in avian species has led to propose that sexual selection drives brain size dimorphism (Madden, 2001, Garamszegi et al., 2005). In primates it has been shown that the largest relative brain sizes are associated with monogamous mating systems, leading to suggest that primate monogamy requires greater social acuity and the ability to manipulate others within the group (Schillaci, 2006). Moreover, among primate species it was also observed that increasing levels of body mass dimorphism are associated with decreasing relative brain size (Schillaci, 2006). In the same line of research, a comparative analysis of brain, testis, and social and mating systems data for more than 300 Chiroptera species showed that species with promiscuous females are associated with smaller brains and larger testes, while species with females exhibiting mate fidelity are associated with significantly larger brains and smaller testicles. Thus supporting a negative evolutionary relationship between investment in testes and the development of brain size due to the metabolic cost of both tissues (Pitnick et al., 2006).

Consistent with these observations, our results reveal that the least dimorphic species have gene families expansions significantly overrepresented in functional categories related to various aspects of brain development.

Testes size relative to body mass has also been proposed to be an indicator of sexual selection, thus, in future analysis it would be interesting to investigate the relationship among gene family size variations and this additional indicator of sexual selection.

The species included in these studies cover most taxa from across the mammalian group. However, it is worth noting that species from the orders Rodentia and Primate are overrepresented (13 and 19 respectively). Both these taxa have been the most studied organisms partly because they were some of the first eukaryotic genomes to be sequenced, and for being the closest related to *Homo sapiens* (similar in terms of physiology, metabolism, behaviour, or disease) (Hodgkinson and Eyre-Walker, 2011). As humans, it is understandable that there is an intrinsic interest in understanding the genomic bases and dispersion in the genome of our own species and related ones driven by curiosity and potential for clinical applications (Rogers and Gibbs, 2014). Rodents are frequently employed to undertake laboratory experiments *in vivo* due to their intrinsic characteristics that make them a suitable model for biomedical research. The distant species platypus was removed from the analyses as it showed to be an outlier in the distribution of set of genes. This mammal belongs to the order Monotremata that originated about 166 Myr ago in the earliest split of this taxa's phylogeny (Bininda-Emonds et al., 2007).

Finally, Chapter 4 marks a depart from the previous two chapters in terms of topic of study as well as differences as well as a change in the type of data examined –from mammals to plants and from genomes to transcriptomes-. In this chapter, the relationship between transcriptome profiles and geographical distribution of 19 samples of the widely distributed sea rocket was examined. This, in order to gain insight into the genetic basis of plant adaptation to changing environments. The emergence and improvement of specialized technology and methods, such as RNA-seq, have facilitated the

study of the molecular processes that shape phenotypic traits in an ever expanding set of species beyond traditional model organisms. The analysis of transcriptomic data has allowed the discovery of molecular mechanisms that enhance our knowledge of the relationship with trait variation and environment (Schoville et al., 2012, Pespeni et al., 2013, Yang et al., 2015, Akman et al., 2016). Regulation of gene expression is an essential strategy for the adaptation to environmental stress conditions (Sham et al., 2015, Akman et al., 2016). It is worth noting that gene expression variations could be influenced by the particular season affecting the same individuals or populations at the time. Therefore the present analysis of the *Cakile maritima* samples stemming from different locations but grown in common and controlled conditions provides information on the intrinsic differential expression. Thus proving that the identified phenotypic variation in response to local conditions is at least partially heritable. Our results suggest a differential impact of the geographical distributions on gene expression in relations with the bioclimatic variables mean temperature, annual precipitation, mean diurnal range, temperature seasonality, mean temperature of wettest quarter and precipitation seasonality. Furthermore, the genes displaying a positive association between gene expression variations and bioclimatic variables revealed a statistically significant overrepresentation in various functional categories related to plant stress responses, including regulation of peptidase activity, sugar metabolism processes, cell redox homeostasis, regulation of gene expression and DNA duplex unwinding. This is true even after correcting for phylogenetic similarity across populations and suggests a role of transcriptional profile changes in adaptation of *C. maritima* to local environmental conditions. Importantly, these genes identified as associated with bioclimatic variables could prove valuable in understanding the response to drought and salt tolerance in economically important crops as *C. maritima* along with model plant *A. thaliana* is a member of the family Brassicaceae (Cruciferae) which also includes broccoli, cauliflower, cabbage, choy sum, rutabaga, turnip and some seeds used in producing canola oil. Overall, these results provide novel insights into the genetic mechanisms of adaptation to changing

environments in plants and critically contribute to our understanding of the vulnerability of plant species to climate change.

Without the past technical and monetary limitations of nucleotide data collection, the addition of more species characterized will permit to increase analyses statistical power. It will also allow representing a wider range of species across taxa and thus the discovery of further genomic differences between organisms. Differences that might play important roles in adaptive evolution, at the gene level as well as large changes in the size of gene families. In general, genomic research is extending to less studied organisms like birds and non-model mammals. For instance, the analysis of an alternative system like birds might improve our understanding on gene family expansion in line with brain features and indexes of sexual selection. Birds are the ideal system to study this potential convergence as they represent an independent evolution of endothermy, high cognitive ability and parental care. There are over 360 bird species with a sequenced genome. This is a far better species coverage than the one used in the studies presented. As shown in all studies, given the number of genes and gene families being tested, sample sizes would need to be significantly higher than the numbers available to us.

In particular, birds could help disentangle potential gene family expansion in line with brain features which could in fact be a by-product of adaptations related to placenta invasiveness shown to be related to some extent with encephalization. This is because birds lay eggs and therefore offspring do not represent the same degree of immune challenge for the mother. Additionally, bird behaviour is much better documented and would open the door to identify genome signatures associated with cognitive ability and sociality aspects.

With regards to indexes of sexual selection, birds represent an ideal system to expand the present study as they have a wider range of mating systems and a more diverse parental care labour division. In addition, birds have a much wider range of plumage pigmentation and a high frequency of pigmentation dimorphism which offers an additional index of sexual selection to size and relative testis size.

From a genomics perspective, studying patterns of gene family variation in birds is very interesting as birds have a much higher degree of gene synteny than mammals. Thus, it is possible that gene duplication events are rarer than in mammals.

Both studies show a significantly higher number of gene families expanding in line with phenotype evolution. However, whether these expansions respond to the need for higher levels of expression of the genes forming a gene family or instead respond to pressures associated with transcript diversity is not known. Both studies would benefit from examining reliable and comparable estimates of both gene expression levels and of alternative isoforms produced by each gene family in several species at extremes of the distribution of the phenotypes tested. It could then be tested whether genes within families associated with each phenotype tend to be associated with higher levels of gene expression and or greater transcript availability.

Regarding the analysis presented in Chapter 4, having identified a set of genes potentially related to climatic variables, it would be interesting to further characterise the genes found to be associated with drought and salinity. This could be done by examining gene enrichment in KEGG pathways (Kanehisa and Goto, 2000). Given the quantity of available gene expression profiles for 19 individuals, it is possible to do a gene network analysis examining whether genes associated with each climatic variable

form part of a single molecular pathway. In addition analyses will benefit by limiting the search space to genes and regulatory elements known to play a role in the relevant cell types involved in the phenotype of interest. For example, root tissues for drought conditions and leafs for salinity tolerance (Brunner et al., 2015, Zeng et al., 2018). Another avenue to confirm the role of the identified candidate genes is to examine mutants for each of these genes in the closely related *Arabidopsis thaliana* species. Plants from these mutant seeds could be grown under normal conditions or under draught and high salt conditions. These experiments would allow to uncover molecular mechanisms associated with local adaptation in plants.

It is also crucial to integrate sequence data with other variables including morphological, functional and ecological to have a wider sense of the relationship between genetics and evolutionary important traits.

Altogether, the diverse studies presented emphasize the utility of the development and application of genomic technologies and distinct type of analysis approaches to a better understanding of the genetic basis of complex phenotypes in evolution. At the same time the results I have presented in this thesis provide novel information that will help to develop new directions in future research.

References

ABOUHEIF, E. & FAIRBAIRN, D. J. 1997. A Comparative Analysis of Allometry for Sexual Size Dimorphism: Assessing Rensch's Rule. *The American Naturalist,* 149**,** 540-562.

AFGAN, E., BAKER, D., VAN DEN BEEK, M., BLANKENBERG, D., BOUVIER, D., CECH, M., CHILTON, J., CLEMENTS, D., CORAOR, N., EBERHARD, C., GRUNING, B., GUERLER, A., HILLMAN-JACKSON, J., VON KUSTER, G., RASCHE, E., SORANZO, N., TURAGA, N., TAYLOR, J., NEKRUTENKO, A. & GOECKS, J. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res,* 44**,** W3-W10.

AIELLO, L. C. & WHEELER, P. 1995. The Expensive-Tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution. *Current Anthropology,* 36**,** 199-221.

AKHTAR, R. S., NESS, J. M. & ROTH, K. A. 2004. Bcl-2 family regulation of neuronal development and neurodegeneration. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research,* 1644**,** 189-203.

AKMAN, M., CARLSON, J. E., HOLSINGER, K. E. & LATIMER, A. M. 2016. Transcriptome sequencing reveals population differentiation in gene expression linked to functional traits and environmental gradients in the South African shrub Protea repens. *New Phytol,* 210**,** 295-309.

ÁLVAREZ, H., SERRANO-MENESES, M. A., REYES-MÁRQUEZ, CORTÉS, J. G. & CÓRDOBA-AGUILAR, A. 2013. Allometry of a sexual trait in relation to diet experience and alternative mating tactics in two rubyspot damselflies (Calopterygidae: Hetaerina). *Biological Journal of the Linnean Society,* 108**,** 521-533.

AMTMANN, A. 2009. Learning from evolution: Thellungiella generates new knowledge on essential and critical components of abiotic stress tolerance in plants. *Mol Plant,* 2**,** 3-12.

ANDERSSON, M. B. 1994. *Sexual selection,* Chichester, Princeton University Press.

ARAÚJO, A., ARRUDA, M. F., ALENCAR, A. I., ALBUQUERQUE, F., NASCIMENTO, M. C. & YAMAMOTO, M. E. 2000. Body Weight of Wild and Captive Common Marmosets (Callithrix jacchus). *International Journal of Primatology,* 21**,** 317-324.

ARBOUR, N., VANDERLUIT, J. L., LE GRAND, J. N., JAHANI-ASL, A., RUZHYNSKY, V. A., CHEUNG, E. C. C., KELLY, M. A., MACKENZIE, A. E., PARK, D. S., OPFERMAN, J. T. & SLACK, R. S. 2008. Mcl-1 is a key regulator of apoptosis during CNS development and after DNA damage. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* 28**,** 6068-6078.

ARCHIBALD, J. D. 2006. Zofia Kielan-Jaworowska, Richard L. Cifelli, and Zhe-Xi Luo, Mammals from the Age of Dinosaurs Origins, Evolution, and Structure Columbia University Press, New York, 2004, 630 pp., $195 (hard cover). ISBN 0-231-11918-6. *Journal of Mammalian Evolution,* 13**,** 147-149.

ARCHIBALD, J. D. & DEUTSCHMAN, D. 2001. Quantitative Analysis of the Timing of the Origin and Diversification of Extant Placental Orders. *Journal of Mammalian Evolution,* 8**,** 107-124.

AZEVEDO, F. A., CARVALHO, L. R., GRINBERG, L. T., FARFEL, J. M., FERRETTI, R. E., LEITE, R. E., JACOB FILHO, W., LENT, R. & HERCULANO-HOUZEL, S. 2009. Equal numbers of

neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol,* 513**,** 532-41.

BAINBRIDGE, D. 2000. Evolution of mammalian pregnancy in the presence of the maternal immune system. *Reviews of Reproduction,* 5**,** 67-74.

BAOTIC, A. & STOEGER, A. S. 2017. Sexual dimorphism in African elephant social rumbles. *PLoS One,* 12**,** e0177411.

BARBOUR, M. G. & RODMAN, J. E. 1970. Saga of the West Coast sea-rockets: Cakile Edentula ssp. Californica and C. Maritima. *Rhodora,* 72**,** 370-386.

BEN HAMED, K., BEN HAMED, I., BOUTEAU, F. & ABDELLY, C. 2016. Insights into the Ecology and the Salt Tolerance of the Halophyte Cakile maritima Using Multidisciplinary Approaches. *In:* KHAN, M. A., OZTURK, M., GUL, B. & AHMED, M. Z. (eds.) *Halophytes for Food Security in Dry Lands.* Elsevier Science.

BENJAMIN, A. M., NICHOLS, M., BURKE, T. W., GINSBURG, G. S. & LUCAS, J. E. 2014. Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC Genomics,* 15**,** 570.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological),* 57**,** 289-300.

BI, S., WANG, Y., GUAN, J., SHENG, X. & MENG, J. 2014. Three new Jurassic euharamiyidan species reinforce early divergence of mammals. *Nature,* 514**,** 579-584.

BININDA-EMONDS, O. & L. GITTLEMAN, J. 2000. Are pinnipeds functionally different from fissiped carnivores? The importance of phylogenetic comparative analyses. *Evolution,* 54**,** 1011-1023.

BININDA-EMONDS, O. R. P., CARDILLO, M., JONES, K. E., MACPHEE, R. D. E., BECK, R. M. D., GRENYER, R., PRICE, S. A., VOS, R. A., GITTLEMAN, J. L. & PURVIS, A. 2007. The delayed rise of present-day mammals. *Nature,* 446**,** 507-512.

BLANCKENHORN, W. U. 2000. The evolution of body size: what keeps organisms small? *Q Rev Biol,* 75**,** 385-407.

BODDY, A. M., MCGOWEN, M. R., SHERWOOD, C. C., GROSSMAN, L. I., GOODMAN, M. & WILDMAN, D. E. 2012. Comparative analysis of encephalization in mammals reveals relaxed constraints on anthropoid primate and cetacean brain scaling. *J Evol Biol,* 25**,** 981-94.

BONDRUP-NIELSEN, S. & IMS, R. A. 1990. Reversed sexual size dimorphism in microtines: Are females larger than males or are males smaller than females? *Evolutionary Ecology,* 4**,** 261-272.

BRESSAN, R. A., ZHANG, C., ZHANG, H., HASEGAWA, P. M., BOHNERT, H. J. & ZHU, J. K. 2001. Learning from the Arabidopsis experience. The next gene search paradigm. *Plant Physiol,* 127**,** 1354-60.

BRUNNER, I., HERZOG, C., DAWES, M. A., AREND, M. & SPERISEN, C. 2015. How tree roots respond to drought. *Frontiers in Plant Science,* 6.

BUDNIK, V. & SALINAS, P. C. 2011. Wnt signaling during synaptic development and plasticity. *Current Opinion in Neurobiology,* 21**,** 151-159.

BUTLER, P. M. 2000. Review of the early allotherian mammals. *Acta Palaeontologica Polonica,* 45**,** 317-342.

CASPARI, E. 1963. Selective Forces in the Evolution of Man. *The American Naturalist,* 97**,** 5-14.

CASTILLO-MORALES, A., MONZON-SANDOVAL, J., DE SOUSA, A. A., URRUTIA, A. O. & GUTIERREZ, H. 2016. Neocortex expansion is linked to size variations in gene families with chemotaxis, cell-cell signalling and immune response functions in mammals. *Open Biol,* 6.

CASTILLO-MORALES, A., MONZON-SANDOVAL, J., URRUTIA, A. O. & GUTIERREZ, H. 2014. Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions. *Proc Biol Sci,* 281**,** 20132428.

CASTILLO, A., MONZON, J., URRUTIA, A. O. & GUTIERREZ, H. 2013. Encephalization level in mammalian species is associated with the expansion of gene families with cell-cell signalling, chemotaxis and immune system-related functions. . *Genome Biol***,** (submitted).

CHARLESWORTH, B. 2012. The Effects of Deleterious Mutations on Evolution at Linked Sites. *Genetics,* 190**,** 5-22.

CHARLTON, B. D., ZHIHE, Z. & SNYDER, R. J. 2009. The information content of giant panda, Ailuropoda melanoleuca, bleats: acoustic cues to sex, age and size. *Animal Behaviour,* 78**,** 893-898.

CHEN, F.-C., CHEN, C.-J., LI, W.-H. & CHUANG, T.-J. 2010. Gene Family Size Conservation Is a Good Indicator of Evolutionary Rates. *Molecular Biology and Evolution,* 27**,** 1750-1758.

CHOTARD, C. & SALECKER, I. 2004. Neurons and glia: team players in axon guidance. *Trends Neurosci,* 27**,** 655-61.

CICCARELLI, D., BALESTRI, M., MARIA PAGNI, A. & FORINO, L. 2010. Morpho-functional adaptations in Cakile maritima Scop. subsp. maritima: Comparation of two different morphological types. *Caryologia -Firenze-,* 63**,** 411-421.

CIFELLI, R. L. & DAVIS, B. M. 2013. Palaeontology: Jurassic fossils and mammalian antiquity. *Nature,* 500**,** 160-1.

CIFELLI, R. L., EBERLE, J. J., LOFGREN, D. L., LILLEGRAVEN, J. A. & CLEMENS, W. A. 2004. Mammalian Biochronology of the Latest Cretaceous. *In:* WOODBURNE, M. O. (ed.) *Late Cretaceous and Cenozoic Mammals of North America.* Columbia University Press.

CIFELLI, R. L. & GORDON, C. L. 2007. Evolutionary biology: Re-crowning mammals. *Nature,* 447**,** 918-920.

CLARK, D. A., MITRA, P. P. & WANG, S. S. 2001. Scalable architecture in mammalian brains. *Nature,* 411**,** 189-93.

CLAUSING, G., VICKERS, K. & KADEREIT, J. W. 2000. Historical biogeography in a linear system: genetic variation of sea rocket (Cakile maritima) and sea holly (Eryngium maritimum) along European coasts. *Mol Ecol,* 9**,** 1823-33.

COHEN, J. 1988. *Statistical power analysis for the behavioral sciences,* Hillsdale, New Jersey, Erlbaum.

CONRAD, B. & ANTONARAKIS, S. E. 2007. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet,* 8**,** 17-35.

CORETEAM 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

COX, J., JACKSON, A. P., BOND, J. & WOODS, C. G. 2006. What primary microcephaly can tell us about brain growth. *Trends in Molecular Medicine,* 12**,** 358-366.

CUMMINGS, S. R. & HULLEY, S. B. 1988a. *Designing clinical research: an epidemiologic approach*, Williams & Wilkins.

CUMMINGS, S. R. & HULLEY, S. B. 1988b. Designing clinical research: an epidemiologic approach. Baltimore: Williams and Wilkins.

CUNNINGHAM, F., ACHUTHAN, P., AKANNI, W., ALLEN, J., AMODE, M. R., ARMEAN, I. M., BENNETT, R., BHAI, J., BILLIS, K. & BODDU, S. 2018. Ensembl 2019. *Nucleic acids research,* 47**,** D745-D751.

DALE, J., DUNN, P. O., FIGUEROLA, J., LISLEVAND, T., SZÉKELY, T. & WHITTINGHAM, L. A. 2007. Sexual selection explains Rensch's rule of allometry for sexual size dimorphism. *Proceedings of the Royal Society B: Biological Sciences,* 274**,** 2971-2979.

DALE, S., LIFJELD, J. T. & ROWE, M. 2015. Commonness and ecology, but not bigger brains, predict urban living in birds. *BMC ecology,* 15**,** 12.

DARWIN, C. 1901. *The descent of man, and selection in relation to sex,* London, London : Murray.

DAVY, A. J., SCOTT, R. & CORDAZZO, C. V. 2006. Biological flora of the British Isles: Cakile maritima Scop. *Journal of Ecology,* 94**,** 695-711.

DE MIGUEL, C. & HENNEBERG, M. 1998. Encephalization of the koala, *Phascolarctos cinereus*. *Australian Mammalogy,* 20**,** 315-320.

DEFELIPE, J. 2011. The evolution of the brain, the human nature of cortical circuits and intellectual creativity. *Frontiers in Neuroanatomy,* 5.

DEMUTH, J. P., BIE, T. D., STAJICH, J. E., CRISTIANINI, N. & HAHN, M. W. 2006. The Evolution of Mammalian Gene Families. *PLoS ONE,* 1**,** e85.

DORMANN, C. F., ELITH, J., BACHER, S., BUCHMANN, C., CARL, G., CARRÉ, G., MARQUÉZ, J. R. G., GRUBER, B., LAFOURCADE, B. & LEITÃO, P. J. 2013a. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography,* 36**,** 27-46.

DORMANN, C. F., ELITH, J., BACHER, S., BUCHMANN, C., CARL, G., CARRÉ, G., MARQUÉZ, J. R. G., GRUBER, B., LAFOURCADE, B., LEITÃO, P. J., MÜNKEMÜLLER, T., MCCLEAN, C., OSBORNE, P. E., REINEKING, B., SCHRÖDER, B., SKIDMORE, A. K., ZURELL, D. & LAUTENBACH, S. 2013b. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography,* 36**,** 27-46.

DOS REIS, M., INOUE, J., HASEGAWA, M., ASHER, R., DONOGHUE, P. & YANG, Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. R. Soc. B-Biol. Sci.,* 279**,** 3491-3500.

DOS REMEDIOS, N., SZEKELY, T., KUPPER, C., LEE, P. L. & KOSZTOLANYI, A. 2015. Ontogenic differences in sexual size dimorphism across four plover populations. *Ibis (Lond 1859),* 157**,** 590-600.

DOS SANTOS, S. E., PORFIRIO, J., DA CUNHA, F. B., MANGER, P. R., TAVARES, W., PESSOA, L., RAGHANTI, M. A., SHERWOOD, C. C. & HERCULANO-HOUZEL, S. 2017. Cellular Scaling Rules for the Brains of Marsupials: Not as "Primitive" as Expected. *Brain Behav Evol,* 89**,** 48-63.

DUNHAM, A. E., MAITNER, B. S., RAZAFINDRATSIMA, O. H., SIMMONS, M. C. & ROY, C. L. 2013. Body size and sexual size dimorphism in primates: influence of climate and net primary productivity. *J Evol Biol,* 26**,** 2312-20.

DURET, L. & MOUCHIROUD, D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular biology and evolution,* 17**,** 68-070.

ELLIOT, M. & CRESPI, B. 2008. Placental invasiveness and brain–body allometry in eutherian mammals. *Journal of evolutionary biology,* 21**,** 1763-1778.

ERWIN, D. H. 2000. Macroevolution is more than repeated rounds of microevolution. *Evolution & Development,* 2**,** 78-84.

EVANS, P. D., ANDERSON, J. R., VALLENDER, E. J., CHOI, S. S. & LAHN, B. T. 2004. Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. *Hum Mol Genet,* 13**,** 1139-45.

EVSYUKOVA, I., PLESTANT, C. & ANTON, E. S. 2013. Integrative Mechanisms of Oriented Neuronal Migration in the Developing Brain. *Annual review of cell and developmental biology,* 29**,** 299-353.

FAIRBAIRN, D. J. 1997. Allometry for Sexual Size Dimorphism: Pattern and Process in the Coevolution of Body Size in Males and Females. *Annual Review of Ecology and Systematics,* 28**,** 659-687.

FAIRBAIRN, D. J., BLANCKENHORN, W. U. & SZEKELY, T. 2007. *Sex, size, and gender roles : evolutionary studies of sexual size dimorphism,* Oxford, Oxford : Oxford University Press.

FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *The American Naturalist,* 125**,** 1-15.

FICK, S. E. & HIJMANS, R. J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology,* 37**,** 4302-4315.

FISH, J. L. & LOCKWOOD, C. A. 2003. Dietary constraints on encephalization in primates. *Am J Phys Anthropol,* 120**,** 171-81.

FLORIO, M., ALBERT, M., TAVERNA, E., NAMBA, T., BRANDL, H., LEWITUS, E., HAFFNER, C., SYKES, A., WONG, F. K., PETERS, J., GUHR, E., KLEMROTH, S., PRUFER, K., KELSO, J., NAUMANN, R., NUSSLEIN, I., DAHL, A., LACHMANN, R., PAABO, S. & HUTTNER, W. B. 2015. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science,* 347**,** 1465-70.

FROSINI, S., LARDICCI, C. & BALESTRI, E. 2012. Global Change and Response of Coastal Dune Plants to the Combined Effects of Increased Sand Accretion (Burial) and Nutrient Availability. *PLOS ONE,* 7**,** e47561.

GABI, M., COLLINS, C. E., WONG, P., TORRES, L. B., KAAS, J. H. & HERCULANO-HOUZEL, S. 2010. Cellular scaling rules for the brains of an extended number of primate species. *Brain Behav Evol,* 76**,** 32-44.

GANDOUR, M., HESSINI, K. & ABDELLY, C. 2008. Understanding the population genetic structure of coastal species (Cakile maritima): seed dispersal and the role of sea currents in determining population structure. *Genet Res (Camb),* 90**,** 167-78.

GARAMSZEGI, L., EENS, M., ERRITZØE, J. & MOLLER, A. 2005. Sperm competition and sexually size dimorphic brains in birds. *Proceedings. Biological sciences / The Royal Society,* 272**,** 159-66.

GARWICZ, M., CHRISTENSSON, M. & PSOUNI, E. 2009. A unifying model for timing of walking onset in humans and other mammals. *Proceedings of the National Academy of Sciences,* 106**,** 21889.

GATTERMANN, R., FRITZSCHE, P., WEINANDY, R. & NEUMANN, K. 2002. Comparative studies of body mass, body measurements and organ weights of wild-derived and laboratory golden hamsters (Mesocricetus auratus). *Lab Anim,* 36**,** 445-54.

GIBSON, K. R., RUMBAUGH, D. & BERAN, M. 2001. Bigger is better: primate brain size in relationship to cognition. *Evolutionary anatomy of the primate cerebral cortex***,** 79-97.

GIGEK, C. O., CHEN, E. S. & SMITH, M. A. C. 2016. Methyl-CpG-Binding Protein (MBD) Family: Epigenomic Read-Outs Functions and Roles in Tumorigenesis and Psychiatric Diseases. *Journal of cellular biochemistry,* 117**,** 29-38.

GILAD, Y., BUSTAMANTE, C. D., LANCET, D. & PÄÄBO, S. 2003. Natural Selection on the Olfactory Receptor Gene Family in Humans and Chimpanzees. *American Journal of Human Genetics,* 73**,** 489-501.

GITTLEMAN, J. L. 1986. Carnivore Brain Size, Behavioral Ecology, and Phylogeny. *Journal of Mammalogy,* 67**,** 23-36.

GONG, Q., LI, P., MA, S., INDU RUPASSARA, S. & BOHNERT, H. J. 2005. Salinity stress adaptation competence in the extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana. *Plant J,* 44**,** 826-39.

GONZALEZ-LAGOS, C., SOL, D. & READER, S. M. 2010. Large-brained mammals live longer. *J Evol Biol,* 23**,** 1064-74.

GONZALEZ-VOYER, A. & VON HARDENBERG, A. 2014. An introduction to phylogenetic path analysis. *Modern phylogenetic comparative methods and their application in evolutionary biology.* Springer.

GOSWAMI, A. 2012. A dating success story: genomes and fossils converge on placental mammal origins. *EvoDevo,* 3**,** 18.

GOULD, S. J. 2002. *The Structure of Evolutionary Theory*, Harvard University Press.

GRAFEN, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences,* 326**,** 119-157.

GRAFEN, A. 1992. The uniqueness of the phylogenetic regression. *Journal of theoretical Biology,* 156**,** 405-423.

GRAHAM, M. H. 2003. Confronting multicollinearity in ecological multiple regression. *Ecology,* 84**,** 2809-2815.

GUTIERREZ, H., CASTILLO, A., MONZON, J. & URRUTIA, A. O. 2011. Protein amino acid composition: a genomic signature of encephalization in mammals. *PLoS One,* 6**,** e27261.

HALBRITTER, A. H., FIOR, S., KELLER, I., BILLETER, R., EDWARDS, P. J., HOLDEREGGER, R., KARRENBERG, S., PLUESS, A. R., WIDMER, A. & ALEXANDER, J. M. 2018. Trait differentiation and adaptation of plants along elevation gradients. *Journal of Evolutionary Biology,* 31**,** 784-800.

HAPPOLD, D. C. D. 1975. The Ecology of Rodents in the Northern Sudan. *In:* PRAKASH, I. & GHOSH, P. K. (eds.) *Rodents in Desert Environments.* Dordrecht: Springer Netherlands.

HÅSTAD, O., VICTORSSON, J. & ÖDEEN, A. 2005. Differences in color vision make passerines less conspicuous in the eyes of their predators. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 6391-6394.

HAUG, H. 1987. Brain sizes, surfaces, and neuronal sizes of the cortex cerebri: a stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *American Journal of Anatomy,* 180**,** 126-142.

HAWKINS, A. & OLSZEWSKI, J. 1957. Glia/nerve cell index for cortex of the whale. *Science,* 126**,** 76-7.

HAYDON, P. G. 2001. Glia: listening and talking to the synapse. *Nat Rev Neurosci,* 2**,** 185-193.

HE, T., FRIEDE, H. & KILIARIDIS, S. 2002. Macroscopic and roentgenographic anatomy of the skull of the ferret (Mustela putorius furo). *Lab Anim,* 36**,** 86-96.

HERCULANO-HOUZEL, S. 2014. The glia/neuron ratio: How it varies uniformly across brain structures and species and what that means for brain physiology and evolution. *Glia,* 62**,** 1377-1391.

HERCULANO-HOUZEL, S., CATANIA, K., MANGER, P. R. & KAAS, J. H. 2015. Mammalian Brains Are Made of These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in the Brain of Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls, and Their Relationship with Body Mass. *Brain Behav Evol,* 86**,** 145-63.

HERCULANO-HOUZEL, S., COLLINS, C. E., WONG, P. & KAAS, J. H. 2007. Cellular scaling rules for primate brains. *Proceedings of the National Academy of Sciences,* 104**,** 3562-3567.

HERCULANO-HOUZEL, S. & LENT, R. 2005. Isotropic fractionator: a simple, rapid method for the quantification of total cell and neuron numbers in the brain. *Journal of Neuroscience,* 25**,** 2518-2521.

HERCULANO-HOUZEL, S., MANGER, P. R. & KAAS, J. H. 2014. Brain scaling in mammalian evolution as a consequence of concerted and mosaic changes in numbers of neurons and average neuronal cell size. *Frontiers in Neuroanatomy,* 8**,** 77.

HERCULANO-HOUZEL, S., RIBEIRO, P., CAMPOS, L., VALOTTA DA SILVA, A., TORRES, L. B., CATANIA, K. C. & KAAS, J. H. 2011. Updated neuronal scaling rules for the brains of Glires (rodents/lagomorphs). *Brain Behav Evol,* 78**,** 302-14.

HERNANDEZ-JIMENEZ, A. & RIOS-CARDENAS, O. 2012. Natural versus sexual selection: predation risk in relation to body size and sexual ornaments in the green swordtail. *Animal Behaviour,* 84**,** 1051-1059.

HIENG, B., UGRINOVIĆ, K., ŠUŠTAR-VOZLIČ, J. & KIDRIČ, M. 2004. Different classes of proteases are involved in the response to drought of Phaseolus vulgaris L. cultivars differing in sensitivity. *Journal of Plant Physiology,* 161**,** 519-530.

HLODAN, O. 2007. Macroevolution: Evolution above the Species Level. *BioScience,* 57**,** 222-225.

HODGKINSON, A. & EYRE-WALKER, A. 2011. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics,* 12**,** 756.

HOFFMANN, A. A. & SGRÒ, C. M. 2011. Climate change and evolutionary adaptation. *Nature,* 470**,** 479.

HOLLAND, P. W., MARLETAZ, F., MAESO, I., DUNWELL, T. L. & PAPS, J. 2017. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philos Trans R Soc Lond B Biol Sci,* 372.

HOSKEN, D. J. & HOUSE, C. M. 2011. Sexual selection. *Current Biology,* 21**,** R62-R65.

HUSAK, J. F., MACEDONIA, J. M., FOX, S. F. & SAUCEDA, R. C. 2006. Predation Cost of Conspicuous Male Coloration in Collared Lizards (Crotaphytus collaris): An Experimental Test Using Clay-Covered Model Lizards. *Ethology,* 112**,** 572-580.

ISLER, K. & VAN SCHAIK, C. P. 2012. Allomaternal care, life history and brain size evolution in mammals. *Journal of Human Evolution,* 63**,** 52-63.

JARDIM-MESSEDER, D., LAMBERT, K., NOCTOR, S., PESTANA, F. M., DE CASTRO LEAL, M. E., BERTELSEN, M. F., ALAGAILI, A. N., MOHAMMAD, O. B., MANGER, P. R. & HERCULANO-HOUZEL, S. 2017. Dogs Have the Most Neurons, Though Not the Largest Brain: Trade-Off between Body Mass and Number of Neurons in the Cerebral Cortex of Large Carnivoran Species. *Front Neuroanat,* 11**,** 118.

JASON A. KAUFMAN, CLAUDE MARCEL HLADIK & PATRICK PASQUET 2003. On the Expensive-Tissue Hypothesis: Independent Support from Highly Encephalized Fish. *Current Anthropology,* 44**,** 705-707.

JEHEE, J. F. M. & MURRE, J. M. J. 2008. The scalable mammalian brain: emergent distributions of glia and neurons. *Biological Cybernetics,* 98**,** 439-445.

JENSEN, C., HAHN, M. E. & DUDEK, B. C. 1979. Chapter I - Introduction: Toward Understanding the Brain—Behavior Relationship. *In:* DUDEK, M. E. H. J. C. (ed.) *Development and Evolution of Brain Size.* Academic Press.

JERISON, H. 2012. *Evolution of the brain and intelligence*, Elsevier.

JERISON, H. J. 1973. *Evolution of the brain and intelligence,* London, Academic Press, Inc.

JERISON, H. J. 1979. Chapter III - The Evolution of Diversity in Brain Size. *In:* DUDEK, M. E. H. J. C. (ed.) *Development and Evolution of Brain Size.* Academic Press.

JERISON, H. J. 2007. What Fossils Tell Us about the Evolution of the Neocortex. *In:* KAAS, J. H. K., L.A. (ed.) *Evolution of Nervous System.* New York and Oxford: Elsevier.

JESSEN, K. R. 2004. Glial cells. *The International Journal of Biochemistry & Cell Biology,* 36**,** 1861-1867.

JIMÉNEZ-ARCOS, V., SANABRIA-URBÁN, S. & CUEVA DEL CASTILLO, R. 2017. The interplay between natural and sexual selection in the evolution of sexual size dimorphism in Sceloporus lizards (Squamata: Phrynosomatidae). *Ecology and Evolution,* 7**,** 905-917.

JOBLING, M. A. A. 2014. *Human evolutionary genetics,* New York, Garland Science.

JONES, K. E., BIELBY, J., CARDILLO, M., FRITZ, S. A., O'DELL, J., ORME, C. D. L., SAFI, K., SECHREST, W., BOAKES, E. H., CARBONE, C., CONNOLLY, C., CUTTS, M. J., FOSTER, J. K., GRENYER, R., HABIB, M., PLASTER, C. A., PRICE, S. A., RIGBY, E. A., RIST, J., TEACHER, A., BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., MACE, G. M. & PURVIS, A. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology,* 90**,** 2648-2648.

JOSEPH, B. & HERMANSON, O. 2010. Molecular control of brain size: Regulators of neural stem cell life, death and beyond. *Experimental Cell Research,* 316**,** 1415-1421.

JUNG, B. P., ZHANG, G., NITSCH, R., TROGADIS, J., NAG, S. & EUBANKS, J. H. 2003. Differential expression of methyl CpG-binding domain containing factor MBD3 in the developing and adult rat brain. *Journal of Neurobiology,* 55**,** 220-232.

KAAS, J. 1993. Evolution of multiple areas and modules within neocortex. *Perspectives on developmental neurobiology,* 1**,** 101-107.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

KAPPELER, P. M., NUNN, C. L., VINING, A. Q. & GOODMAN, S. M. 2019. Evolutionary dynamics of sexual size dimorphism in non-volant mammals following their independent colonization of Madagascar. *Scientific Reports,* 9**,** 1454.

KAZU, R. S., MALDONADO, J., MOTA, B., MANGER, P. R. & HERCULANO-HOUZEL, S. 2014. Cellular scaling rules for the brain of Artiodactyla include a highly folded cortex with few neurons. *Frontiers in Neuroanatomy,* 8**,** 128.

KEENEY, J. G., DAVIS, J. M., SIEGENTHALER, J., POST, M. D., NIELSEN, B. S., HOPKINS, W. D. & SIKELA, J. M. 2015. DUF1220 protein domains drive proliferation in human neural stem cells and are associated with increased cortical volume in anthropoid primates. *Brain Struct Funct,* 220**,** 3053-60.

KEMP, T. S. 2005. *The origin and evolution of mammals*, Oxford University Press Oxford.

KERSEY, P. J., ALLEN, J. E., ARMEAN, I., BODDU, S., BOLT, B. J., CARVALHO-SILVA, D., CHRISTENSEN, M., DAVIS, P., FALIN, L. J., GRABMUELLER, C., HUMPHREY, J., KERHORNOU, A., KHOBOVA, J., ARANGANATHAN, N. K., LANGRIDGE, N., LOWY, E., MCDOWALL, M. D., MAHESWARI, U., NUHN, M., ONG, C. K., OVERDUIN, B., PAULINI, M., PEDRO, H., PERRY, E., SPUDICH, G., TAPANARI, E., WALTS, B., WILLIAMS, G., TELLO–RUIZ, M., STEIN, J., WEI, S., WARE, D., BOLSER, D. M., HOWE, K. L., KULESHA, E., LAWSON, D., MASLEN, G. & STAINES, D. M. 2015. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research,* 44**,** D574-D580.

KEUNEN, E., PESHEV, D., VANGRONSVELD, J., VAN DEN ENDE, W. & CUYPERS, A. 2013. Plant sugars are crucial players in the oxidative challenge during abiotic stress: extending the traditional concept. *Plant Cell Environ,* 36**,** 1242-55.

KINGSLEY, M. C. S. 1979. Fitting the von Bertalanffy growth equation to polar bear age–weight data. *Canadian Journal of Zoology,* 57**,** 1020-1025.

KINGSOLVER, J. G. & PFENNIG, D. W. 2004. Individual-level selection as a cause of Cope's rule of phyletic size increase. *Evolution,* 58**,** 1608-12.

KINSELLA, R. J., KAHARI, A., HAIDER, S., ZAMORA, J., PROCTOR, G., SPUDICH, G., ALMEIDA-KING, J., STAINES, D., DERWENT, P., KERHORNOU, A., KERSEY, P. & FLICEK, P. 2011.

Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford),* 2011**,** bar030.

KLEIMAN, D. G. 1977. Monogamy in mammals. *Q Rev Biol,* 52**,** 39-69.

KLENOVŠEK, T. & KRYŠTUFEK, B. 2013. An ontogenetic perspective on the study of sexual dimorphism, phylogenetic variability, and allometry of the skull of European ground squirrel, Spermophilus citellus (Linnaeus, 1766). *Zoomorphology,* 132**,** 433-445.

KNOCK, E., PEREIRA, J., LOMBARD, P. D., DIMOND, A., LEAFORD, D., LIVESEY, F. J. & HENDRICH, B. 2015. The methyl binding domain 3/nucleosome remodelling and deacetylase complex regulates neural cell fate determination and terminal differentiation in the cerebral cortex. *Neural Development,* 10**,** 13.

KOENIG, D. & WEIGEL, D. 2015. Beyond the thale: comparative genomics and genetics of Arabidopsis relatives. *Nat Rev Genet,* 16**,** 285-98.

KOIWA, H., LI, F., MCCULLY, M. G., MENDOZA, I., KOIZUMI, N., MANABE, Y., NAKAGAWA, Y., ZHU, J., RUS, A., PARDO, J. M., BRESSAN, R. A. & HASEGAWA, P. M. 2003. The STT3a subunit isoform of the Arabidopsis oligosaccharyltransferase controls adaptive responses to salt/osmotic stress. *The Plant cell,* 15**,** 2273-2284.

KRAUSE, D. W., HOFFMANN, S., WIBLE, J. R., KIRK, E. C., SCHULTZ, J. A., VON KOENIGSWALD, W., GROENKE, J. R., ROSSIE, J. B., O/'CONNOR, P. M., SEIFFERT, E. R., DUMONT, E. R., HOLLOWAY, W. L., ROGERS, R. R., RAHANTARISOA, L. J., KEMP, A. D. & ANDRIAMIALISON, H. 2014. First cranial remains of a gondwanatherian mammal reveal remarkable mosaicism. *Nature,* 515**,** 512-517.

KUMA, K.-I., IWABE, N. & MIYATA, T. 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. *Molecular biology and evolution,* 12**,** 123-130.

KUMAR, S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet,* 6**,** 654-662.

KUMAR, S., STECHER, G., SULESKI, M. & HEDGES, S. B. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol,* 34**,** 1812-1819.

KURTA, A. & KUNZ, T. H. 1988. Roosting Metabolic Rate and Body Temperature of Male Little Brown Bats (Myotis lucifugus) in Summer. *Journal of Mammalogy,* 69**,** 645-651.

KUZAWA, C. W., CHUGANI, H. T., GROSSMAN, L. I., LIPOVICH, L., MUZIK, O., HOF, P. R., WILDMAN, D. E., SHERWOOD, C. C., LEONARD, W. R. & LANGE, N. 2014. Metabolic costs and evolutionary implications of human brain development. *Proc Natl Acad Sci U S A,* 111**,** 13010-5.

LANGFELDER, P. & HORVATH, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* 9**,** 559.

LAUWERS, F., CASSOT, F., LAUWERS-CANCES, V., PUWANARAJAH, P. & DUVERNOY, H. 2008. Morphometry of the human cerebral cortex microcirculation: general characteristics and space-related profiles. *Neuroimage,* 39**,** 936-48.

LAW, C. W., ALHAMDOOSH, M., SU, S., DONG, X., TIAN, L., SMYTH, G. K. & RITCHIE, M. E. 2016. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research,* 5**,** ISCB Comm J-1408.

LEBLOND, C. S., HEINRICH, J., DELORME, R., PROEPPER, C., BETANCUR, C., HUGUET, G., KONYUKH, M., CHASTE, P., EY, E., RASTAM, M., ANCKARSÄTER, H., NYGREN, G., GILLBERG, I. C., MELKE, J., TORO, R., REGNAULT, B., FAUCHEREAU, F., MERCATI, O., LEMIÈRE, N., SKUSE, D., POOT, M., HOLT, R., MONACO, A. P., JÄRVELÄ, I., KANTOJÄRVI, K., VANHALA, R., CURRAN, S., COLLIER, D. A., BOLTON, P., CHIOCCHETTI, A., KLAUCK, S. M., POUSTKA, F., FREITAG, C. M., WALTES, R., KOPP, M., DUKETIS, E., BACCHELLI, E., MINOPOLI, F., RUTA, L., BATTAGLIA, A., MAZZONE, L., MAESTRINI, E., SEQUEIRA, A. F., OLIVEIRA, B., VICENTE, A., OLIVEIRA, G., PINTO, D., SCHERER, S. W., ZELENIKA, D., DELEPINE, M., LATHROP, M., BONNEAU, D., GUINCHAT, V., DEVILLARD, F., ASSOULINE, B., MOUREN, M.-C., LEBOYER, M., GILLBERG, C., BOECKERS, T. M. & BOURGERON, T. 2012. Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. *PLoS genetics,* 8**,** e1002521-e1002521.

LEE, B. H., LEE, H., XIONG, L. & ZHU, J. K. 2002. A mitochondrial complex I defect impairs cold-regulated nuclear gene expression. *Plant Cell,* 14**,** 1235-51.

LEE, T.-H., GUO, H., WANG, X., KIM, C. & PATERSON, A. H. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics,* 15**,** 162.

LEFEBVRE, L. 2012. Primate encephalization. *Prog Brain Res,* 195**,** 393-412.

LEIMU, R. & FISCHER, M. 2008. A Meta-Analysis of Local Adaptation in Plants. *PLOS ONE,* 3**,** e4010.

LEONARD, W. R., SNODGRASS, J. J. & ROBERTSON, M. L. 2007. Effects of brain evolution on human nutrition and metabolism. *Annu Rev Nutr,* 27**,** 311-27.

LERNER, K. L. 2013. *Gale Encyclopedia of Science (3rd, 4th, and 5th editions)*, Cengage Gale.

LEVINTON, J. S. 1983. Stasis in Progress: The Empirical Basis of Macroevolution. *Annual Review of Ecology and Systematics,* 14**,** 103-137.

LEVINTON, J. S. 2001. *Genetics, Paleontology, and Macroevolution,* Cambridge, Cambridge : Cambridge University Press.

LEWITUS, E., HOF, P. R. & SHERWOOD, C. C. 2012. Phylogenetic comparison of neuron and glia densities in the primary visual cortex and hippocampus of carnivores and primates. *Evolution,* 66**,** 2551-63.

LI, J. & JI, L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity,* 95**,** 221.

LIM, H., CHO, M. H., JEON, J. S., BHOO, S. H., KWON, Y. K. & HAHN, T. R. 2009. Altered expression of pyrophosphate: fructose-6-phosphate 1-phosphotransferase affects the growth of transgenic Arabidopsis plants. *Mol Cells,* 27**,** 641-9.

LINDENFORS, P., L GITTLEMAN, J. & JONES, K. 2007. Sexual size dimorphism in mammals. Oxford: Oxford University Press.

LINDENFORS, P. & S.TULLBERG, B. 2011. 2 - Evolutionary Aspects of Aggression: The Importance of Sexual Selection. *In:* HUBER, R., BANNASCH, D. L. & BRENNAN, P. (eds.) *Advances in Genetics.* Academic Press.

LIU, X., HAN, S., WANG, Z., GELERNTER, J. & YANG, B. Z. 2013. Variant callers for next-generation sequencing data: a comparison study. *PLoS One,* 8**,** e75619.

LOBRÉAUX, S. & MIQUEL, C. 2019a. Identification of Arabis alpina genomic regions associated with climatic variables along an elevation gradient through whole genome scan. *Genomics*.

LOBRÉAUX, S. & MIQUEL, C. 2019b. Identification of Arabis alpina genomic regions associated with climatic variables along an elevation gradient through whole genome scan. *Genomics*, doi.org/10.1016/j.ygeno.2019.05.008.

LOPEZ-MAURY, L., MARGUERAT, S. & BAHLER, J. 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet,* 9**,** 583-93.

LOVE, M. I., ANDERS, S., KIM, V. & HUBER, W. 2015. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research,* 4**,** 1070-1070.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15**,** 550.

LUNTER, G. & GOODSON, M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research,* 21**,** 936-939.

LUO, Z.-X. 2007. Transformation and diversification in early mammal evolution. *Nature,* 450**,** 1011-1019.

LYNCH, M. & CONERY, J. S. 2000. The evolutionary fate and consequences of duplicate genes. *Science,* 290**,** 1151-5.

M. MCDONOUGH, C. 2000. *Social Organization of Nine-banded Armadillos (Dasypus novemcinctus) in a Riparian Habitat*.

MACDONALD, D. W. & SILLERO-ZUBIRI, C. 2004. *The biology and conservation of wild canids,* Oxford, Oxford University Press.

MADDEN, J. 2001. Sex, bowers and brains. *Proceedings. Biological sciences,* 268**,** 833-838.

MAGADUM, S., BANERJEE, U., MURUGAN, P., GANGAPUR, D. & RAVIKESAVAN, R. 2013. Gene duplication as a major force in evolution. *J Genet,* 92**,** 155-61.

MARIANI, J., SIMONINI, M. V., PALEJEV, D., TOMASINI, L., COPPOLA, G., SZEKELY, A. M., HORVATH, T. L. & VACCARINO, F. M. 2012. Modeling human cortical development in vitro using induced pluripotent stem cells. *Proceedings of the National Academy of Sciences,* 109**,** 12770.

MARÍN, O., VALIENTE, M., GE, X. & TSAI, L.-H. 2010. Guiding Neuronal Cell Migrations. *Cold Spring Harbor Perspectives in Biology,* 2**,** a001834.

MARTIN, A. R., COSTA, H. A., LAPPALAINEN, T., HENN, B. M., KIDD, J. M., YEE, M.-C., GRUBERT, F., CANN, H. M., SNYDER, M. & MONTGOMERY, S. B. 2014. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS genetics,* 10**,** e1004549.

MATEOS-APARICIO, P. & RODRÍGUEZ-MORENO, A. 2019. The Impact of Studying Brain Plasticity. *Frontiers in cellular neuroscience,* 13**,** 66-66.

MCINTYRE, L. M., LOPIANO, K. K., MORSE, A. M., AMIN, V., OBERG, A. L., YOUNG, L. J. & NUZHDIN, S. V. 2011. RNA-seq: technical variability and sampling. *BMC Genomics,* 12**,** 293.

MCLAIN, D. K. 1993. Cope's Rules, Sexual Selection, and the Loss of Ecological Plasticity. *Oikos,* 68**,** 490-500.

MCNAB, B. K. & EISENBERG, J. F. 1989. Brain Size and Its Relation to the Rate of Metabolism in Mammals. *The American Naturalist,* 133**,** 157-167.

MCPHERSON, F. J. & CHENOWETH, P. J. 2012. Mammalian sexual dimorphism. *Animal Reproduction Science,* 131**,** 109-122.

MEGHA, S., BASU, U. & KAV, N. N. V. 2018. Regulation of low temperature stress in plants by microRNAs. *Plant, Cell & Environment,* 41**,** 1-15.

MENG, J. 2014. Mesozoic mammals of China: implications for phylogeny and early evolution of mammals. *National Science Review,* 1**,** 521-542.

MENON, S. & GUPTON, S. L. 2016. Building Blocks of Functioning Brain: Cytoskeletal Dynamics in Neuronal Development. *International review of cell and molecular biology,* 322**,** 183-245.

MEREDITH, R. W., JANECKA, J. E., GATESY, J., RYDER, O. A., FISHER, C. A., TEELING, E. C., GOODBLA, A., EIZIRIK, E., SIMAO, T. L., STADLER, T., RABOSKY, D. L., HONEYCUTT, R. L., FLYNN, J. J., INGRAM, C. M., STEINER, C., WILLIAMS, T. L., ROBINSON, T. J., BURK-HERRICK, A., WESTERMAN, M., AYOUB, N. A., SPRINGER, M. S. & MURPHY, W. J. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science,* 334**,** 521-4.

MILLER, G., SHULAEV, V. & MITTLER, R. 2008. Reactive oxygen signaling and abiotic stress. *Physiol Plant,* 133**,** 481-9.

MILLER, J. A., DING, S.-L., SUNKIN, S. M., SMITH, K. A., NG, L., SZAFER, A., EBBERT, A., RILEY, Z. L., ROYALL, J. J., AIONA, K., ARNOLD, J. M., BENNET, C., BERTAGNOLLI, D., BROUNER, K., BUTLER, S., CALDEJON, S., CAREY, A., CUHACIYAN, C., DALLEY, R. A., DEE, N., DOLBEARE, T. A., FACER, B. A. C., FENG, D., FLISS, T. P., GEE, G., GOLDY, J., GOURLEY, L., GREGOR, B. W., GU, G., HOWARD, R. E., JOCHIM, J. M., KUAN, C. L., LAU, C., LEE, C.-K., LEE, F., LEMON, T. A., LESNAR, P., MCMURRAY, B., MASTAN, N., MOSQUEDA, N., NALUAI-CECCHINI, T., NGO, N.-K., NYHUS, J., OLDRE, A., OLSON, E., PARENTE, J., PARKER, P. D., PARRY, S. E., STEVENS, A., PLETIKOS, M., REDING, M., ROLL, K., SANDMAN, D., SARREAL, M., SHAPOURI, S., SHAPOVALOVA, N. V., SHEN, E. H., SJOQUIST, N., SLAUGHTERBECK, C. R., SMITH, M., SODT, A. J., WILLIAMS, D., ZÖLLEI, L., FISCHL, B., GERSTEIN, M. B., GESCHWIND, D. H., GLASS, I. A., HAWRYLYCZ, M. J., HEVNER, R. F., HUANG, H., JONES, A. R., KNOWLES, J. A., LEVITT, P., PHILLIPS, J. W., SESTAN, N., WOHNOUTKA, P., DANG, C., BERNARD, A., HOHMANN, J. G. & LEIN, E. S. 2014. Transcriptional landscape of the prenatal human brain. *Nature,* 508**,** 199-206.

MONTGOMERY, S. H., CAPELLINI, I., VENDITTI, C., BARTON, R. A. & MUNDY, N. I. 2011. Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. *Mol Biol Evol,* 28**,** 625-38.

MONTIEL, J. F., KAUNE, H. & MALIQUEO, M. 2013. Maternal-fetal unit interactions and eutherian neocortical development and evolution. *Frontiers in neuroanatomy,* 7**,** 22.

MONZÓN-SANDOVAL, J., CASTILLO-MORALES, A., CRAMPTON, S., MCKELVEY, L., NOLAN, A., O'KEEFFE, G. & GUTIERREZ, H. 2015. Modular and coordinated expression of immune system regulatory and signaling components in the developing and adult nervous system. *Frontiers in Cellular Neuroscience,* 9**,** 337.

MORALES, A. C. 2015. *Genomic Signatures of Neurodegeneration and the Evolution of Mammalian Brain.* PhD, University of Bath.

MORAND, S. & RICKLEFS, R. E. 2005. Genome size is not related to life-history traits in primates. *Genome,* 48**,** 273-8.

MORGANS, C. L., COOKE, G. M. & ORD, T. J. 2014. How populations differentiate despite gene flow: sexual and natural selection drive phenotypic divergence within a land fish, the Pacific leaping blenny. *BMC Evolutionary Biology,* 14**,** 97-97.

MYSTERUD, A. 2000. The relationship between ecological segregation and sexual body size dimorphism in large herbivores. *Oecologia,* 124**,** 40-54.

NAKAGAWA, S. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology,* 15**,** 1044-1045.

NAVARRETE, A., VAN SCHAIK, C. P. & ISLER, K. 2011. Energetics and the evolution of human brain size. *Nature,* 480**,** 91-3.

NEDERGAARD, M., RANSOM, B. & GOLDMAN, S. A. 2003. New roles for astrocytes: redefining the functional architecture of the brain. *Trends Neurosci,* 26**,** 523-30.

NENGOMASHA, E. M., PEARSON, R. A. & SMITH, T. 2016. The donkey as a draught power resource in smallholder farming in semi-arid western Zimbabwe: 1. Live weight and food and water requirements. *Animal Science,* 69**,** 297-304.

NIELSEN, R., PAUL, J. S., ALBRECHTSEN, A. & SONG, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet,* 12**,** 443-451.

NIIMURA, Y. & NEI, M. 2005. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene,* 346**,** 13-21.

NOONAN, M. J., JOHNSON, P. J., KITCHENER, A. C., HARRINGTON, L. A., NEWMAN, C. & MACDONALD, D. W. 2016. Sexual size dimorphism in musteloids: An anomalous allometric pattern is explained by feeding ecology. *Ecology and Evolution,* 6**,** 8495-8501.

NORTHCUTT, R. G. 2002. Understanding Vertebrate Brain Evolution. *Integrative and Comparative Biology,* 42**,** 743-756.

NOWAK, R. M. & DICKMAN, C. R. 2005. *Walker's Marsupials of the World*, Johns Hopkins University Press.

NUNES, A. C., MATHIAS, M. L. & CRESPO, A. M. 2001. Morphological and haematological parameters in the Algerian mouse (Mus spretus) inhabiting an area contaminated with heavy metals. *Environ Pollut,* 113**,** 87-93.

NUNNEY, L. 2015. Adapting to a Changing Environment: Modeling the Interaction of Directional Selection and Plasticity. *Journal of Heredity,* 107**,** 15-24.

O'MARA, M. T., GORDON, A. D., CATLETT, K. K., TERRANOVA, C. J. & SCHWARTZ, G. T. 2012. Growth and the development of sexual size dimorphism in lorises and galagos. *Am J Phys Anthropol,* 147**,** 11-20.

OPFERMAN, J. T. & KOTHARI, A. 2017. Anti-apoptotic BCL-2 family members in development. *Cell Death And Differentiation,* 25**,** 37.

PARADIS, E. & SCHLIEP, K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics,* 35**,** 526-528.

PARADIS, E. & SCHLIEP, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics,* 35**,** 526-528.

PARK, S. M., JANG, H. J. & LEE, J. H. 2019. Roles of Primary Cilia in the Developing Brain. *Front Cell Neurosci,* 13**,** 218.

PÉREZ-BARBERÍA, F. J., GORDON, I. J. & PAGEL, M. 2002. The Origins of Sexual Dimorphism in Body Size in Ungulates. *Evolution,* 56**,** 1276-1285.

PESPENI, M. H., BARNEY, B. T. & PALUMBI, S. R. 2013. Differences in the regulation of growth and biomineralization genes revealed through long-term common-garden acclimation and experimental genomics in the purple sea urchin. *Evolution,* 67**,** 1901-14.

PINHEIRO, J., BATES, D., DEBROY, S. & SARKAR, D. 2018. R Core Team (2018). nlme: linear and nonlinear mixed effects models. R package version 3.1-137.

PINTO, M., JEPSEN, K. J., TERRANOVA, C. J. & BUFFENSTEIN, R. 2010. Lack of sexual dimorphism in femora of the eusocial and hypogonadic naked mole-rat: a novel animal model for the study of delayed puberty on the skeletal system. *Bone,* 46**,** 112-20.

PIROOZNIA, M., KRAMER, M., PARLA, J., GOES, F. S., POTASH, J. B., MCCOMBIE, W. R. & ZANDI, P. P. 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics,* 8**,** 14.

PITNICK, S., JONES, K. E. & WILKINSON, G. S. 2006. Mating system and brain size in bats. *Proc Biol Sci,* 273**,** 719-24.

PREUSS, T. M. 2012. Human brain evolution: From gene discovery to phenotype discovery. *Proceedings of the National Academy of Sciences,* 109**,** 10709-10716.

RAIKWAR, S., SRIVASTAVA, V. K., GILL, S. S., TUTEJA, R. & TUTEJA, N. 2015. Emerging Importance of Helicases in Plant Stress Tolerance: Characterization of Oryza sativa Repair Helicase XPB2 Promoter and Its Functional Validation in Tobacco under Multiple Stresses. *Frontiers in plant science,* 6**,** 1094-1094.

READ, A. J., WELLS, R. S., HOHN, A. A. & SCOTT, M. D. 1993. Patterns of growth in wild bottlenose dolphins, Tursiops truncatus. *Journal of Zoology,* 231**,** 107-123.

RESEARCH, W. 2016a. Amniote Life History Database. Wolfram Data Repository.

RESEARCH, W. 2016b. Amniote Life History Database. Wolfram Data Repository.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England),* 26**,** 139-140.

RODRIGUES, R. J., MARQUES, J. M. & CUNHA, R. A. 2018. Purinergic signalling and brain development. *Semin Cell Dev Biol*.

ROGERS, J. & GIBBS, R. A. 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature reviews. Genetics,* 15**,** 347-359.

ROSENTHAL, R., ROSNOW, R. L. & RUBIN, D. B. 2000. *Contrasts and effect sizes in behavioral research: A correlational approach*, Cambridge University Press.

ROTH, G. & DICKE, U. 2005. Evolution of the brain and intelligence. *Trends in cognitive sciences,* 9**,** 250-257.

ROTH, G. & WULLIMANN, M. F. 2001. *Brain Evolution and Cognition*, Wiley.

RUDOY, A. & RIBERA, I. 2017. Evolution of sexual dimorphism and Rensch's rule in the beetle genus Limnebius (Hydraenidae): is sexual selection opportunistic? *PeerJ,* 5**,** e3060.

SACHER, G. A. & STAFFELDT, E. F. 1974. Relation of Gestation Time to Brain Weight for Placental Mammals: Implications for the Theory of Vertebrate Growth. *The American Naturalist,* 108**,** 593-615.

SALA, C., VICIDOMINI, C., BIGI, I., MOSSA, A. & VERPELLI, C. 2015. Shank synaptic scaffold proteins: keys to understanding the pathogenesis of autism and other synaptic disorders. *Journal of Neurochemistry,* 135**,** 849-858.

SAN MAURO, D. & AGORRETA, A. 2010. Molecular systematics: A synthesis of the common methods and the state of knowledge. *Cell Mol Biol Lett,* 15**,** 311-41.

SASSONE, J., MARASCHI, A., SASSONE, F., SILANI, V. & CIAMMOLA, A. 2013. Defining the role of the Bcl-2 family proteins in Huntington's disease. *Cell Death &Amp; Disease,* 4**,** e772.

SCHILLACI, M. A. 2006. Sexual selection and the evolution of brain size in primates. *PloS one,* 1**,** e62-e62.

SCHOENEMANN, P. T. 2013. Hominid Brain Evolution. *A Companion to Paleoanthropology.* Blackwell Publishing Ltd.

SCHOVILLE, S. D., BARRETO, F. S., MOY, G. W., WOLFF, A. & BURTON, R. S. 2012. Investigating the molecular basis of local adaptation to thermal stress: population differences in gene expression across the transcriptome of the copepod Tigriopus californicus. *BMC Evol Biol,* 12**,** 170.

SCHRAGO, C. G., MELLO, B. & SOARES, A. E. R. 2013. Combining fossil and molecular data to date the diversification of New World Primates. *Journal of Evolutionary Biology,* 26**,** 2438-2446.

SEIRADAKE, E., DEL TORO, D., NAGEL, D., COP, F., HÄRTL, R., RUFF, T., SEYIT-BREMER, G., HARLOS, K., BORDER, E. C., ACKER-PALMER, A., JONES, E. Y. & KLEIN, R. 2014. FLRT structure: balancing repulsion and cell adhesion in cortical and vascular development. *Neuron,* 84**,** 370-385.

SERRANO-MENESES, M. A., CORDOBA-AGUILAR, A., AZPILICUETA-AMORIN, M., GONZALEZ-SORIANO, E. & SZEKELY, T. 2008. Sexual selection, sexual size dimorphism and Rensch's rule in Odonata. *J Evol Biol,* 21**,** 1259-73.

SHAM, A., MOUSTAFA, K., AL-AMERI, S., AL-AZZAWI, A., IRATNI, R. & ABUQAMAR, S. 2015. Identification of Arabidopsis Candidate Genes in Response to Biotic and Abiotic Stresses Using Comparative Microarrays. *PLoS ONE,* 10**,** e0125666.

SHAO, H. B., GUO, Q. J., CHU, L. Y., ZHAO, X. N., SU, Z. L., HU, Y. C. & CHENG, J. F. 2007. Understanding molecular mechanism of higher plant plasticity under abiotic stress. *Colloids Surf B Biointerfaces,* 54**,** 37-45.

SHERWOOD, C. C., STIMPSON, C. D., RAGHANTI, M. A., WILDMAN, D. E., UDDIN, M., GROSSMAN, L. I., GOODMAN, M., REDMOND, J. C., BONAR, C. J., ERWIN, J. M. & HOF, P. R. 2006. Evolution of increased glia–neuron ratios in the human frontal cortex. *Proceedings of the National Academy of Sciences,* 103**,** 13606-13611.

SHULTZ, S. & DUNBAR, R. 2010. Encephalization is not a universal macroevolutionary phenomenon in mammals but is associated with sociality. *Proceedings of the National Academy of Sciences*.

SMITH, J. M. 1998. *Evolutionary genetics,* Oxford, Oxford : Oxford University Press.

SMITH, R. J. & JUNGERS, W. L. 1997. Body mass in comparative primatology. *J Hum Evol,* 32**,** 523-59.

SOCIÉTÉ DES PRODUITS NESTLÉ S.A. 2019. *Purina* [Online]. Available: https://www.purina.com/dogs/dog-breeds/boxer [Accessed May 2018].

SOSHNIKOVA, N., DEWAELE, R., JANVIER, P., KRUMLAUF, R. & DUBOULE, D. 2013. Duplications of hox gene clusters and the emergence of vertebrates. *Developmental Biology,* 378**,** 194-199.

STEPHAN, H., NELSON, J. E. & FRAHM, H. D. 1981. Brain size comparison in Chiroptera. *Journal of Zoological Systematics and Evolutionary Research,* 19**,** 195-222.

STRACHAN, T. & READ, A. 2010. *Human Molecular Genetics*, Taylor & Francis Group.

STRIEDTER, G. F. 2005. *Principles of brain evolution,* Sunderland, Mass., Sunderland, Mass. : Sinauer Associates.

SUAREZ, R. & MPODOZIS, J. 2009. Heterogeneities of size and sexual dimorphism between the subdomains of the lateral-innervated accessory olfactory bulb (AOB) of Octodon degus (Rodentia: Hystricognathi). *Behav Brain Res,* 198**,** 306-12.

SUN, D., ZHOU, X., YU, H.-L., HE, X.-X., GUO, W.-X., XIONG, W.-C. & ZHU, X.-J. 2017. Regulation of neural stem cell proliferation and differentiation by Kinesin family member 2a. *PLOS ONE,* 12**,** e0179047.

SUZUKI, N. & MITTLER, R. 2006. Reactive oxygen species and temperature stresses: A delicate balance between signaling and destruction. *Physiologia Plantarum,* 126**,** 45-51.

SYLVESTER, J. B., RICH, C. A., LOH, Y.-H. E., VAN STAADEN, M. J., FRASER, G. J. & STREELMAN, J. T. 2010. Brain diversity evolves via differences in patterning. *Proceedings of the National Academy of Sciences,* 107**,** 9718-9723.

TACUTU, R., THORNTON, D., JOHNSON, E., BUDOVSKY, A., BARARDO, D., CRAIG, T., DIANA, E., LEHMANN, G., TOREN, D., WANG, J., FRAIFELD, V. E. & DE MAGALHAES, J. P. 2018. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res,* 46**,** D1083-D1090.

TAKEUCHI, O. & AKIRA, S. 2010. Pattern recognition receptors and inflammation. *Cell,* 140**,** 805-20.

TAYLOR, J. S. & RAES, J. 2004. Duplication and Divergence: The Evolution of New Genes and Old Ideas. *Annual Review of Genetics,* 38**,** 615-643.

TAYLOR, J. S. & RAES, J. 2005. Chapter 5 - Small-Scale Gene Duplications. *In:* GREGORY, T. R. (ed.) *The Evolution of the Genome.* Burlington: Academic Press.

TEMPLETON, A. R. 2006. *Population Genetics and Microevolutionary Theory*, Wiley.

TICKLE, C. & URRUTIA, A. O. 2017. Perspectives on the history of evo-devo and the contemporary research landscape in the genomics era. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences,* 372**,** 20150473.

TOWER, D. B. & YOUNG, O. M. 1973. The activities of butyrylcholinesterase and carbonic anhydrase, the rate of anaerobic glycolysis, and the question of a constant density of glial cells in cerebral cortices of various mammalian species from mouse to whale. *J Neurochem,* 20**,** 269-78.

TURNER, T. L., BOURNE, E. C., VON WETTBERG, E. J., HU, T. T. & NUZHDIN, S. V. 2010. Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nat Genet,* 42**,** 260-263.

TUTEJA, N., SINGH, S. & TUTEJA, R. 2012. Helicases in Improving Abiotic Stress Tolerance in Crop Plants. *Improving Crop Resistance to Abiotic Stress*, 435-449.

UYEDA, J. C., HANSEN, T. F., ARNOLD, S. J. & PIENAAR, J. 2011. The million-year wait for macroevolutionary bursts. *Proceedings of the National Academy of Sciences,* 108, 15908-15913.

VAN DER BIJL, W. 2018. phylopath: Easy phylogenetic path analysis in R. *PeerJ,* 6, e4718.

VIERSTRA, R. D. 1996. Proteolysis in plants: mechanisms and functions. *In:* FILIPOWICZ, W. & HOHN, T. (eds.) *Post-Transcriptional Control of Gene Expression in Plants.* Dordrecht: Springer Netherlands.

WAGNER, G. P., AMEMIYA, C. & RUDDLE, F. 2003. Hox cluster duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences of the United States of America,* 100, 14603-14606.

WAN, J.-X., ZHU, X.-F., WANG, Y.-Q., LIU, L.-Y., ZHANG, B.-C., LI, G.-X., ZHOU, Y.-H. & ZHENG, S.-J. 2018. Xyloglucan Fucosylation Modulates Arabidopsis Cell Wall Hemicellulose Aluminium binding Capacity. *Scientific Reports,* 8, 428.

WANG, H.-Y., CHIEN, H.-C., OSADA, N., HASHIMOTO, K., SUGANO, S., GOJOBORI, T., CHOU, C.-K., TSAI, S.-F., WU, C.-I. & SHEN, C.-K. J. 2006. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS biology,* 5, e13.

WANG, S. S. H., SHULTZ, J. R., BURISH, M. J., HARRISON, K. H., HOF, P. R., TOWNS, L. C., WAGERS, M. W. & WYATT, K. D. 2008. Shaping of white matter composition by biophysical scaling constraints. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* 28, 4047-4056.

WANG, X., MCCOY, P. A., RODRIGUIZ, R. M., PAN, Y., JE, H. S., ROBERTS, A. C., KIM, C. J., BERRIOS, J., COLVIN, J. S., BOUSQUET-MOORE, D., LORENZO, I., WU, G., WEINBERG, R. J., EHLERS, M. D., PHILPOT, B. D., BEAUDET, A. L., WETSEL, W. C. & JIANG, Y.-H. 2011. Synaptic dysfunction and abnormal behaviors in mice lacking major isoforms of Shank3. *Human molecular genetics,* 20, 3093-3108.

WARTON, D. I., WRIGHT, I. J., FALSTER, D. S. & WESTOBY, M. 2006. Bivariate line-fitting methods for allometry. *Biol Rev Camb Philos Soc,* 81, 259-91.

WECKERLY, F. W. 1998. Sexual-Size Dimorphism: Influence of Mass and Mating Systems in the Most Dimorphic Mammals. *Journal of Mammalogy,* 79, 33-52.

WEIL, A. 2014. Mammalian evolution: A beast of the southern wild. *Nature,* 515, 495-496.

WEISBECKER, V. & GOSWAMI, A. 2010. Brain size, life history, and metabolism at the marsupial/placental dichotomy. *Proceedings of the National Academy of Sciences,* 107, 16216.

WILKINSON, G. S., BREDEN, F., MANK, J. E., RITCHIE, M. G., HIGGINSON, A. D., RADWAN, J., JAQUIERY, J., SALZBURGER, W., ARRIERO, E., BARRIBEAU, S. M., PHILLIPS, P. C., RENN, S. C. & ROWE, L. 2015. The locus of sexual selection: moving sexual selection studies into the post-genomics era. *J Evol Biol,* 28, 739-55.

WILLIAMS, S. E., SHOO, L. P., ISAAC, J. L., HOFFMANN, A. A. & LANGHAM, G. 2008. Towards an Integrated Framework for Assessing the Vulnerability of Species to Climate Change. *PLOS Biology,* 6, e325.

WOLFF, J. O. 1985. Comparative population ecology of Peromyscus leucopus and Peromyscus maniculatus. *Canadian Journal of Zoology,* 63, 1548-1555.

WONG, C. E., LI, Y., LABBE, A., GUEVARA, D., NUIN, P., WHITTY, B., DIAZ, C., GOLDING, G. B., GRAY, G. R., WERETILNYK, E. A., GRIFFITH, M. & MOFFATT, B. A. 2006. Transcriptional profiling implicates novel interactions between abiotic stress and hormonal responses in Thellungiella, a close relative of Arabidopsis. *Plant Physiol,* 140**,** 1437-50.

WONG, C. E., LI, Y., WHITTY, B. R., DIAZ-CAMINO, C., AKHTER, S. R., BRANDLE, J. E., GOLDING, G. B., WERETILNYK, E. A., MOFFATT, B. A. & GRIFFITH, M. 2005. Expressed sequence tags from the Yukon ecotype of Thellungiella reveal that gene expression in response to cold, drought and salinity shows little overlap. *Plant Mol Biol,* 58**,** 561-74.

WU, H. J., ZHANG, Z., WANG, J. Y., OH, D. H., DASSANAYAKE, M., LIU, B., HUANG, Q., SUN, H. X., XIA, R., WU, Y., WANG, Y. N., YANG, Z., LIU, Y., ZHANG, W., ZHANG, H., CHU, J., YAN, C., FANG, S., ZHANG, J., WANG, Y., ZHANG, F., WANG, G., LEE, S. Y., CHEESEMAN, J. M., YANG, B., LI, B., MIN, J., YANG, L., WANG, J., CHU, C., CHEN, S. Y., BOHNERT, H. J., ZHU, J. K., WANG, X. J. & XIE, Q. 2012. Insights into salt tolerance from the genome of Thellungiella salsuginea. *Proc Natl Acad Sci U S A,* 109**,** 12219-24.

YAMANE, A., DOI, T. & ONO, Y. 1996. Mating behaviors, courtship rank and mating success of male feral cat (Felis catus). *Journal of Ethology,* 14**,** 35-44.

YANG, Y., LI, X., KONG, X., MA, L., HU, X. & YANG, Y. 2015. Transcriptome analysis reveals diversified adaptation of Stipa purpurea along a drought gradient on the Tibetan Plateau. *Funct Integr Genomics,* 15**,** 295-307.

YEO, G., HOLSTE, D., KREIMAN, G. & BURGE, C. B. 2004. Variation in alternative splicing across human tissues. *Genome biology,* 5**,** R74.

ZENG, F., SHABALA, S., MAKSIMOVIĆ, J. D., MAKSIMOVIĆ, V., BONALES-ALATORRE, E., SHABALA, L., YU, M., ZHANG, G. & ŽIVANOVIĆ, B. D. 2018. Revealing mechanisms of salinity tissue tolerance in succulent halophytes: A case study for Carpobrotus rossi. *Plant, Cell & Environment,* 41**,** 2654-2667.

ZHANG, B. & HORVATH, S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol,* 4**,** Article17.

ZHANG, J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution,* 18**,** 292-298.

ZHANG, L., HU, X., MIAO, X., CHEN, X., NAN, S. & FU, H. 2016. Genome-Scale Transcriptome Analysis of the Desert Shrub Artemisia sphaerocephala. *PLOS ONE,* 11**,** e0154300.

ZHAO, X., UEBA, T., CHRISTIE, B. R., BARKHO, B., MCCONNELL, M. J., NAKASHIMA, K., LEIN, E. S., EADIE, B. D., WILLHOITE, A. R., MUOTRI, A. R., SUMMERS, R. G., CHUN, J., LEE, K.-F. & GAGE, F. H. 2003. Mice lacking methyl-CpG binding protein 1 have deficits in adult neurogenesis and hippocampal function. *Proceedings of the National Academy of Sciences,* 100**,** 6777.

ZHENG, X., BI, S., WANG, X. & MENG, J. 2013. A new arboreal haramiyid shows the diversity of crown mammals in the Jurassic period. *Nature,* 500**,** 199-202.

ZHOU, C.-F., WU, S., MARTIN, T. & LUO, Z.-X. 2013. A Jurassic mammaliaform and the earliest mammalian evolutionary adaptations. *Nature,* 500**,** 163-167.